ANALYSIS AND MAPPING OF RNA POLYMERASE II TERMINATION FACTORS

by

Paul Schaughency

A dissertation submitted to Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy.

Baltimore, Maryland

October, 2014

ABSTRACT

Termination of transcription is an important, but poorly understood, cellular process. RNA polymerase II (Pol II) terminates coding transcripts and non-coding transcripts (ncRNAs) through two distinct mechanisms in yeast. In the non-poly(A) dependent pathway, Nrd1 and Nab3 bind to their recognition sites on the nascent RNA of ncRNAs to cause termination. This is done through an unknown mechanism, but is known to involve the RNA-DNA helicase Sen1. In our first study, we have mapped Pol II distribution before and after depletion of different termination factors. This was done, by depleting each factor from the nucleus and then using PAR-CLIP to assess the position of Pol II. Nrd1 depletion leads to widespread runaway transcription at non-poly(A) terminators. In contrast, depletion of Sen1 or Ysh1, an endonuclease involved in the poly(A) dependent pathway, does not lead to elevated readthrough transcription. These differences allowed us to map all the non-poly(A) termination regions throughout the yeast genome. Our second study follows the human homolog of Nrd1, RBM16. RBM16 shares a common protein domain architecture and a conserved CTD interacting domain with Nrd1. Although it is unclear whether there is a similar non-poly(A) termination pathway in human cells, we show that there are at least similarities between the classes of RNAs that are bound to both Nrd1 and RBM16. The final study focuses on the unique methodologies used to analyze PAR-CLIP datasets. It describes computer programs and bioinformatic pipelines written to provide answers to basic questions based on PAR-CLIP data. The functions created to answer these questions could be applicable to a wide variety of data. The studies presented here provide a foundation so that fundamental mechanisms of termination can be elucidated.

Thesis Advisor: Jeffry Corden, PhD.

Thesis Reader: Randall Reed, PhD.

PREFACE

I would like to thank my thesis advisor Jeffry Corden. Without his support, none of what I write below would have happened. I would also like to thank the past and present members of the Corden and Cormack labs, my thesis committee members (Randall Reed, Jon Lorsch, Rachel Green, and Brendan Cormack), my family, my friends, and the BCMB administration staff for encouragement and support during these past years.

Years later, as I look back upon that fresh recruit into the BCMB program, I hardly recognize that person. I am thankful I have had the opportunity to learned many uncountable things over these past years. In truth, though, I could not have done it without the support of many people. I know I have left someone out in this brief preface, but be assured that I do appreciated all the help people have given me over the years here at Hopkins. As I end this chapter in my life, I realized I now have a solid foundation to build a career in science.

Paul Schaughency

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION

RNA Polymerase II (Pol II) is responsible for all messenger RNA (mRNA) and some non-coding RNA (ncRNA) transcription in eukaryotes. The proteins involved, and the time at which they bind, can define the cycle of transcription. Recruitment and initiation of Pol II on the DNA template is the first step. This is closely followed by elongation of the nascent RNA and, eventual termination of the transcript. The cycle of transcription has been extensively studied and reviewed (Barrero and Malik, 2013; Shandilya and Roberts, 2012; Svetlov and Nudler, 2013). Proper termination of Pol II is essential to maintaining a pool of free enzyme and to prevent Pol II from interfering with downstream genes (Kuehner et al., 2011; Mischo and Proudfoot, 2013). This could happen through collision with oncoming Pol II or dislodging transcription factors on the downstream DNA template (Castelnuovo et al., 2013; Cloutier et al., 2013; Shearwin et al., 2005). In yeast, Pol II termination has been shown to provide a regulatory purpose through premature termination of certain genes (Creamer, 2011; Kuehner and Brow, 2008; Thiebaut et al., 2008).

The C-terminal domain (CTD) of Pol II contains a heptad repeat that acts a scaffold for protein recruitment during the transcriptional cycle (Buratowski, 2009). The consensus sequence (YSPTSPS) of this CTD is conserved from yeast to mammals. Each repeat of the CTD is modified during the transcription cycle to reflect the state of Pol II transcription (Chapman et al., 2007; Zhang and Corden, 1991). This "CTD Code" is a set of phosphorylation and other post-translational modifications that alter which sets of proteins are bound to the scaffold and, ultimately, influence Pol II transcription

(Buratowski, 2009; Carroll et al., 2007; Komarnitsky et al., 2000; Nedea et al., 2003). The most well understood part of the CTD code is phosphorylation of serine residues within the heptad repeat. After initiation of Pol II on the DNA template and early elongation, there is an abundance of phosphorylation on the fifth serine (Ser5) in the tandem repeats. At the 3' end of the template, there is more phosphorylation of the second serine (Ser2). These patterns allow the CTD to be a dynamic scaffold for many different co-transcriptional machines (Corden, 2013; Hsin and Manley, 2012; Napolitano et al., 2014).

Pol II pervasively transcribes many ncRNAs throughout the yeast genome (Berretta and Morillon, 2009; Jacquier, 2009; Jensen et al., 2013). These ncRNAS can be categorized into three subclasses: Cryptic Unstable Transcripts (CUTs) that are readily degraded by the nuclear exosome (Arigo et al., 2006; Davis and Ares, 2006; Gudipati et al., 2012; Neil et al., 2009; Thiebaut et al., 2006; Wyers et al., 2005), Stable uncharacterized transcripts (SUTs) that do not rely on the nuclear exosome to be observed (Xu et al., 2009) and Xrn1-dependent transcripts (XUTs) that are only found in the absence of the nuclease Xrn1 (van Dijk et al., 2011).

In yeast, termination of mRNAs and some SUTs happen through the cleavage and polyadenylation (pA) machinery. Loosely conserved sequences at the 3' end of the transcript act as signal for binding of the CF1 complex, which leads to the recruitment of the pA machinery and termination of the transcript (Kuehner et al., 2011; Mischo and Proudfoot, 2013). The mechanism of this action has yet to be elucidated. snoRNA and other ncRNA termination happens through the Nrd1-Nab3-Sen1 (NNS) pathway. Unlike pA-mediated termination, the NNS pathway binds early in transcription when Ser5

phosphorylation of the CTD is high (Chinchilla et al., 2012). Similarly to pA-termination, Nrd1 and Nab3 recognized RNA sequences elements downstream of certain ncRNAs (Carroll et al., 2007; Carroll et al., 2004; Creamer, 2011; Porrua et al., 2012; Steinmetz et al., 2001; Steinmetz et al., 2006; Wlotzka et al., 2011). This leads to recruitment of the DNA/RNA helicase Sen1 as well as the nuclear exosome (Conrad et al., 2000; Steinmetz and Brow, 1996, 1998; Steinmetz et al., 2001; Ursic et al., 2004). How these factors work together to direct termination and turnover is not well understood.

Two possible mechanisms for transcriptional termination have been proposed. In the first, the exposed uncapped 5' end of the nascent transcript, after cleavage from the poly-A site, recruits a 5' -> 3' exonuclease. In yeast, this protein is Rat1 while in metazoans the protein is Xrn2 (Kim et al., 2004; West et al., 2004). In this model, Pol II is already paused downstream of the poly-A site. The exonuclease degrades the nascent RNA until it reaches the stalled Pol II and apparently this facilitates release. This process is called the "torpedo" model of transcription termination. The other commonly accepted model of transcription termination is referred to as the allosteric model

(Logan et al., 1987). In this model, it is not the outside action of an exonuclease that causes the release but the pausing and binding of certain factors that induce an allosteric change in Pol II, facilitating its release from the DNA template on non-poly(A) transcripts. Pcf11, a member of the pA complex, has been shown to dismantle the elongation complex in vitro (Zhang and Gilmour, 2006; Zhang et al., 2007) and Nrd1 may act similarly. The DNA/RNA helicase Sen1 has been proposed to act like the bacterial termination factor rho and pry off the elongating Pol II (Brow, 2011; Kim et al., 1999; Kuehner et al., 2011; Steinmetz and Brow, 1996; Ursic et al., 2004). As there is

evidence for both models of transcription termination, a mixture of both of these models may be correct.

We have previously used a method called Photoactivatable Ribonucleoside-Enhanced Crosslinking and Immunoprecipitation (PAR-CLIP) to map the binding sites of Nrd1, Nab3, Sen1, and Rpb2 throughout the transcriptome (Creamer, 2011; Jamonnak, 2011). PAR-CLIP is a method to crosslink RNA bound to specific proteins following the incorporation of 4-thiouridine (4TU) or 4-thiouracil (4SU) into the transcriptome (Hafner et al., 2010). 4SU is used in yeast because of it can be readily fed into the yeast nucleotide salvage pathway and incorporated into RNA. 4TU is substituted for 4SU in mammalian cells because these cells do not have the correct uracil phosphoribotransferase enzyme for 4SU to be incorporated (Miller et al., 2009).

Nrd1, Nab3 and Sen1 binding sites were mapped genome-wide as the first step to understand more about the NNS pathway (Creamer, 2011). Nrd1 and Nab3 binding sites were found, not only at the 3' end of snoRNA genes, but also at the 5' end of some mRNA genes and antisense to poorly transcribed genes. This allowed us to formulate a model by which the NNS pathway binds early in transcription and monitors the nascent RNA for binding sites. In this way, the NNS complex may act as a checkpoint for proper elongation. In the case of antisense transcripts and some mRNAs, the complex would find it's binding site and terminate the transcript (Creamer, 2011).

In the first part of these studies, we have used PAR-CLIP to map Pol II globally after mutation of both termination pathways in yeast. Although much was learned from the binding sites of Nrd1, Nab3, and Sen1 in vivo from purifying their bound RNAs, their

global effect on transcription remained untested. This was studied by introducing a new system of depleting tagged factors from the nucleus in a time dependent manner termed Anchor Away. The anchor away scheme uses the binding of rapamycin to induce export of a tagged protein out of the nucleus (Haruki et al., 2008). Using anchor away and Pol II PAR-CLIP, we could visualize the affect of termination factor mutation on termination globally. The Cramer lab (Munich) published this idea recently, but they used 4TU-seq to identify the messages affected by Nrd1 depletion (Schulz et al., 2013). Although, both methods use 4TU, we show that they overestimated the effect of Nrd1 mutation because their method was confounded by RNA turnover. We describe a clear definition of termination regions based on the amount of readthrough caused by termination factor depletion (Schaughency, 2014).

Poly(A)-mediated termination is fairly conserved in higher eukaryotes, but it is unclear whether a pathway similar to the NNS modifies ncRNA transcription. snoRNAs are mostly contained within introns of mammalian genes, but the mechanism of snoRNA maturation still requires the nuclear exosome (Berndt et al., 2012; Hirose et al., 2003). Sen1 has a conserved homolog called Senataxin in mammalian cells (Chen et al., 2006). It is unclear whether the putative helicase Senataxin retains a function in processing ncRNAs . Although, mutations in senataxin have been implicated in the cause of certain rare forms of ALS (Anheim et al., 2009; Bassuk et al., 2007; Chen et al., 2006). Nrd1 does have two homologs identified through their similarity in protein domain architecture and conservation of the CTD interacting domain (CID) (Yuryev et al., 1996). These are RBM16 and SFRS15, but little else about their function within the nucleus is known.

In the second part of these studies, we looked at the binding of RBM16 throughout the mammalian transcriptome. Much like the experiments in mapping the yeast binding sites in vivo, this is the first step in formulating a model for the action of these proteins. We find that RBM16 binds to some protein coding genes, ncRNAs, and snoRNAs. These three categories of bound RNAs seem to be conserved from yeast Nrd1. This may suggest a similar mechanism of action. Further mutation studies could illuminate RBM16's function within the nucleus.

For the final part of these studies, we looked at the analyses of PAR-CLIP datasets. PAR-CLIP requires a unique set of programs to accommodate its analysis problems. There is a T -> C transition when the reverse transcriptase makes the cDNA from the crosslinked RNA (Hafner et al., 2010). Although useful in finding exact crosslinking sites, it poses problems for correct alignment of these sequences to the genome. This was solved, by allowing mismatches when mapping these datasets. Another problem was the normalization PAR-CLIP datasets within an experiment. Without normalization of the reads, every dataset would not be comparable to each other. The solution was to multiply every read by the amount of reads mapped per 10^7 reads. Unique functions and programs were derived in order to answer basic questions about all these PAR-CLIP datasets. Taken together, these studies have extended the knowledge of termination factors in yeast and mammalian cells.

REFERENCES

Anheim, M., Monga, B., Fleury, M., Charles, P., Barbot, C., Salih, M., Delaunoy, J.P., Fritsch, M., Arning, L., Synofzik, M., *et al.* (2009). Ataxia with oculomotor apraxia type 2: clinical, biological and genotype/phenotype correlation study of a cohort of 90 patients. Brain *132*, 2688-2698.

Arigo, J.T., Eyler, D.E., Carroll, K.L., and Corden, J.L. (2006). Termination of cryptic unstable transcripts is directed by yeast RNA-binding proteins Nrd1 and Nab3. Mol Cell *23*, 841-851.

Barrero, M.J., and Malik, S. (2013). The RNA polymerase II transcriptional machinery and its epigenetic context. Sub-cellular biochemistry *61*, 237-259.

Bassuk, A.G., Chen, Y.Z., Batish, S.D., Nagan, N., Opal, P., Chance, P.F., and Bennett, C.L. (2007). In cis autosomal dominant mutation of Senataxin associated with tremor/ataxia syndrome. Neurogenetics *8*, 45-49.

Berndt, H., Harnisch, C., Rammelt, C., Stohr, N., Zirkel, A., Dohm, J.C., Himmelbauer, H., Tavanez, J.P., Huttelmaier, S., and Wahle, E. (2012). Maturation of mammalian H/ACA box snoRNAs: PAPD5-dependent adenylation and PARN-dependent trimming. RNA *18*, 958-972.

Berretta, J., and Morillon, A. (2009). Pervasive transcription constitutes a new level of eukaryotic genome regulation. EMBO Rep *10*, 973-982.

Brow, D.A. (2011). Sen-sing RNA Terminators. Mol Cell *42*, 717-718.

Buratowski, S. (2009). Progression through the RNA polymerase II CTD cycle. Mol Cell *36*, 541-546.

Carroll, K.L., Ghirlando, R., Ames, J.M., and Corden, J.L. (2007). Interaction of yeast RNA-binding proteins Nrd1 and Nab3 with RNA polymerase II terminator elements. RNA *13*, 361-373.

Carroll, K.L., Pradhan, D.A., Granek, J.A., Clarke, N.D., and Corden, J.L. (2004). Identification of cis elements directing termination of yeast nonpolyadenylated snoRNA transcripts. Mol Cell Biol *24*, 6241-6252.

Castelnuovo, M., Rahman, S., Guffanti, E., Infantino, V., Stutz, F., and Zenklusen, D. (2013). Bimodal expression of PHO84 is modulated by early termination of antisense transcription. Nat Struct Mol Biol *20*, 851-858.

Chapman, R.D., Heidemann, M., Albert, T.K., Mailhammer, R., Flatley, A., Meisterernst, M., Kremmer, E., and Eick, D. (2007). Transcribing RNA polymerase II is phosphorylated at CTD residue serine-7. Science *318*, 1780-1782.

Chen, Y.Z., Hashemi, S.H., Anderson, S.K., Huang, Y., Moreira, M.C., Lynch, D.R., Glass, I.A., Chance, P.F., and Bennett, C.L. (2006). Senataxin, the yeast Sen1p orthologue: characterization of a unique protein in which recessive mutations cause ataxia and dominant mutations cause motor neuron disease. Neurobiol Dis *23*, 97-108.

Chinchilla, K., Rodriguez-Molina, J.B., Ursic, D., Finkel, J.S., Ansari, A.Z., and Culbertson, M.R. (2012). Interactions of Sen1, Nrd1, and Nab3 with multiple phosphorylated forms of the Rpb1 C-terminal domain in Saccharomyces cerevisiae. Eukaryot Cell *11*, 417-429.

Cloutier, S.C., Wang, S., Ma, W.K., Petell, C.J., and Tran, E.J. (2013). Long noncoding RNAs promote transcriptional poising of inducible genes. PLoS Biol *11*, e1001715.

Conrad, N.K., Wilson, S.M., Steinmetz, E.J., Patturajan, M., Brow, D.A., Swanson, M.S., and Corden, J.L. (2000). A yeast heterogeneous nuclear ribonucleoprotein complex associated with RNA polymerase II. Genetics *154*, 557-571.

Corden, J.L. (2013). RNA polymerase II C-terminal domain: Tethering transcription to transcript and template. Chemical reviews *113*, 8423-8455.

Creamer, T.J., Darby, M.M., Jamonnak, N., Schaughency, P, Hao, H., Wheelan, S.J., Corden, J.L. (2011). Transcriptome-wide binding sites for componenets of the Saccharomyces cerevisiae non-poly(A) termination pathway: Nrd1, Nab3 and Sen1. PLoS Genetics *7*, e1002329.

Davis, C.A., and Ares, M., Jr. (2006). Accumulation of unstable promoter-associated transcripts upon loss of the nuclear exosome subunit Rrp6p in Saccharomyces cerevisiae. Proc Natl Acad Sci U S A *103*, 3262-3267.

Gudipati, R.K., Xu, Z., Lebreton, A., Seraphin, B., Steinmetz, L.M., Jacquier, A., and Libri, D. (2012). Extensive degradation of RNA precursors by the exosome in wild-type cells. Mol Cell *48*, 409-421.

Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano, M., Jr., Jungkamp, A.C., Munschauer, M.*, et al.* (2010). Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. Cell *141*, 129-141.

Haruki, H., Nishikawa, J., and Laemmli, U.K. (2008). The anchor-away technique: rapid, conditional establishment of yeast mutant phenotypes. Mol Cell *31*, 925-932.

Hirose, T., Shu, M.D., and Steitz, J.A. (2003). Splicing-dependent and -independent modes of assembly for intron-encoded box C/D snoRNPs in mammalian cells. Mol Cell *12*, 113-123.

Hsin, J.P., and Manley, J.L. (2012). The RNA polymerase II CTD coordinates transcription and RNA processing. Genes Dev *26*, 2119-2137.

Jacquier, A. (2009). The complex eukaryotic transcriptome: unexpected pervasive transcription and novel small RNAs. Nat Rev Genet *10*, 833-844.

Jamonnak, N., Creamer, T., Darby, M., Schaughency, P. Wheelan, S., Corden, J. (2011). Yeast Nrd1, Nab3 and Sen1 transcriptome-wide binding maps suggest multiple roles in post-transcriptional RNA processing. RNA *17*, 2011-2025.

Jensen, T.H., Jacquier, A., and Libri, D. (2013). Dealing with pervasive transcription. Mol Cell *52*, 473-484.

Kim, H.D., Choe, J., and Seo, Y.S. (1999). The sen1(+) gene of Schizosaccharomyces pombe, a homologue of budding yeast SEN1, encodes an RNA and DNA helicase. Biochemistry *38*, 14697-14710.

Kim, M., Krogan, N.J., Vasiljeva, L., Rando, O.J., Nedea, E., Greenblatt, J.F., and Buratowski, S. (2004). The yeast Rat1 exonuclease promotes transcription termination by RNA polymerase II. Nature *432*, 517-522.

Komarnitsky, P., Cho, E.J., and Buratowski, S. (2000). Different phosphorylated forms of RNA polymerase II and associated mRNA processing factors during transcription. Genes Dev *14*, 2452-2460.

Kuehner, J.N., and Brow, D.A. (2008). Regulation of a eukaryotic gene by GTP-dependent start site selection and transcription attenuation. Mol Cell *31*, 201-211.

Kuehner, J.N., Pearson, E.L., and Moore, C. (2011). Unravelling the means to an end: RNA polymerase II transcription termination. Nat Rev Mol Cell Biol *12*, 283-294.

Logan, J., Falck-Pedersen, E., Darnell, J.E., Jr., and Shenk, T. (1987). A poly(A) addition site and a downstream termination region are required for efficient cessation of transcription by RNA polymerase II in the mouse beta maj-globin gene. Proc Natl Acad Sci U S A *84*, 8306-8310.

Miller, M.R., Robinson, K.J., Cleary, M.D., and Doe, C.Q. (2009). TU-tagging: cell type-specific RNA isolation from intact complex tissues. Nature methods *6*, 439-441.

Mischo, H.E., and Proudfoot, N.J. (2013). Disengaging polymerase: terminating RNA polymerase II transcription in budding yeast. Biochim Biophys Acta *1829*, 174-185.

Napolitano, G., Lania, L., and Majello, B. (2014). RNA polymerase II CTD modifications: how many tales from a single tail. Journal of cellular physiology *229*, 538-544.

Nedea, E., He, X., Kim, M., Pootoolal, J., Zhong, G., Canadien, V., Hughes, T., Buratowski, S., Moore, C.L., and Greenblatt, J. (2003). Organization and function of APT, a subcomplex of the yeast cleavage and polyadenylation factor involved in the formation of mRNA and small nucleolar RNA 3'-ends. J Biol Chem *278*, 33000-33010.

Neil, H., Malabat, C., d'Aubenton-Carafa, Y., Xu, Z., Steinmetz, L.M., and Jacquier, A. (2009). Widespread bidirectional promoters are the major source of cryptic transcripts in yeast. Nature *457*, 1038-1042.

Porrua, O., Hobor, F., Boulay, J., Kubicek, K., D'Aubenton-Carafa, Y., Gudipati, R.K., Stefl, R., and Libri, D. (2012). In vivo SELEX reveals novel sequence and structural determinants of Nrd1-Nab3-Sen1-dependent transcription termination. EMBO J *31*, 3935-3948.

Schaughency, P., Merran, J., Corden J.L. (2014). Genome-wide mapping of Yeast RNA Polymerase II Termination PLoS Genetics *In Print*.

Schulz, D., Schwalb, B., Kiesel, A., Baejen, C., Torkler, P., Gagneur, J., Soeding, J., and Cramer, P. (2013). Transcriptome surveillance by selective termination of noncoding RNA synthesis. Cell *155*, 1075-1087.

Shandilya, J., and Roberts, S.G. (2012). The transcription cycle in eukaryotes: from productive initiation to RNA polymerase II recycling. Biochim Biophys Acta *1819*, 391-400.

Shearwin, K.E., Callen, B.P., and Egan, J.B. (2005). Transcriptional interference--a crash course. Trends Genet *21*, 339-345.

Steinmetz, E.J., and Brow, D.A. (1996). Repression of gene expression by an exogenous sequence element acting in concert with a heterogeneous nuclear ribonucleoprotein-like protein, Nrd1, and the putative helicase Sen1. Mol Cell Biol *16*, 6993-7003.

Steinmetz, E.J., and Brow, D.A. (1998). Control of pre-mRNA accumulation by the essential yeast protein Nrd1 requires high-affinity transcript binding and a domain implicated in RNA polymerase II association. Proc Natl Acad Sci U S A *95*, 6699-6704.

Steinmetz, E.J., Conrad, N.K., Brow, D.A., and Corden, J.L. (2001). RNA-binding protein Nrd1 directs poly(A)-independent 3'-end formation of RNA polymerase II transcripts. Nature *413*, 327-331.

Steinmetz, E.J., Ng, S.B., Cloute, J.P., and Brow, D.A. (2006). cis- and trans-Acting determinants of transcription termination by yeast RNA polymerase II. Mol Cell Biol *26*, 2688-2696.

Svetlov, V., and Nudler, E. (2013). Basic mechanism of transcription by RNA polymerase II. Biochim Biophys Acta *1829*, 20-28.

Thiebaut, M., Colin, J., Neil, H., Jacquier, A., Seraphin, B., Lacroute, F., and Libri, D. (2008). Futile cycle of transcription initiation and termination modulates the response to nucleotide shortage in S. cerevisiae. Mol Cell *31*, 671-682.

Thiebaut, M., Kisseleva-Romanova, E., Rougemaille, M., Boulay, J., and Libri, D. (2006). Transcription termination and nuclear degradation of cryptic unstable transcripts: a role for the nrd1-nab3 pathway in genome surveillance. Mol Cell *23*, 853-864.

Ursic, D., Chinchilla, K., Finkel, J.S., and Culbertson, M.R. (2004). Multiple protein/protein and protein/RNA interactions suggest roles for yeast DNA/RNA helicase Sen1p in transcription, transcription-coupled DNA repair and RNA processing. Nucleic Acids Res *32*, 2441-2452.

van Dijk, E.L., Chen, C.L., d'Aubenton-Carafa, Y., Gourvennec, S., Kwapisz, M., Roche, V., Bertrand, C., Silvain, M., Legoix-Ne, P., Loeillet, S.*, et al.* (2011). XUTs are a class of Xrn1-sensitive antisense regulatory non-coding RNA in yeast. Nature *475*, 114-117.

West, S., Gromak, N., and Proudfoot, N.J. (2004). Human 5' --> 3' exonuclease Xrn2 promotes transcription termination at co-transcriptional cleavage sites. Nature *432*, 522-525.

Wlotzka, W., Kudla, G., Granneman, S., and Tollervey, D. (2011). The nuclear RNA polymerase II surveillance system targets polymerase III transcripts. EMBO J *30*, 1790-1803.

Wyers, F., Rougemaille, M., Badis, G., Rousselle, J.C., Dufour, M.E., Boulay, J., Regnault, B., Devaux, F., Namane, A., Seraphin, B.*, et al.* (2005). Cryptic pol II transcripts are degraded by a nuclear quality control pathway involving a new poly(A) polymerase. Cell *121*, 725-737.

Xu, Z., Wei, W., Gagneur, J., Perocchi, F., Clauder-Munster, S., Camblong, J., Guffanti, E., Stutz, F., Huber, W., and Steinmetz, L.M. (2009). Bidirectional promoters generate pervasive transcription in yeast. Nature *457*, 1033-1037.

Yuryev, A., Patturajan, M., Litingtung, Y., Joshi, R.V., Gentile, C., Gebara, M., and Corden, J.L. (1996). The C-terminal domain of the largest subunit of RNA polymerase II interacts with a novel set of serine/arginine-rich proteins. Proc Natl Acad Sci U S A *93*, 6975-6980.

Zhang, J., and Corden, J.L. (1991). Phosphorylation causes a conformational change in the carboxyl-terminal domain of the mouse RNA polymerase II largest subunit. J Biol Chem *266*, 2297-2302.

Zhang, Z., and Gilmour, D.S. (2006). Pcf11 is a termination factor in Drosophila that dismantles the elongation complex by bridging the CTD of RNA polymerase II to the nascent transcript. Mol Cell *21*, 65-74.

Zhang, Z., Klatt, A., Henderson, A.J., and Gilmour, D.S. (2007). Transcription termination factor Pcf11 limits the processivity of Pol II on an HIV provirus to repress gene expression. Genes Dev *21*, 1609-1614.

Part 1: Genome-wide mapping of Yeast RNA Polymerase II Termination

Paul Schaughency, Jonathan Merran and Jeffry L. Corden

This work has been published in PLoS Genetics and its inclusion in this dissertation is in accordance with the Creative Commons Licensing Agreement.

ABSTRACT

Yeast RNA polymerase II (Pol II) terminates transcription of coding transcripts through the polyadenylation (pA) pathway and non-coding transcripts through the non-polyadenylation (non-pA) pathway. We have used PAR-CLIP to map the position of Pol II genome-wide in living yeast cells after depletion of components of either the pA or non-pA termination complexes. We show here that Ysh1, responsible for cleavage at the pA site, is required for efficient removal of Pol II from the template. Depletion of Ysh1 from the nucleus does not, however, lead to readthrough transcription. In contrast, depletion of the termination factor Nrd1 leads to widespread runaway elongation of non-pA transcripts. Depletion of Sen1 also leads to readthrough at non-pA terminators, but in contrast to Nrd1 this readthrough is less processive or more susceptible to pausing. The data presented here provide delineation of in vivo Pol II termination regions and highlight differences in the sequences that signal termination of different classes of non-pA transcripts.

AUTHOR SUMMARY

Transcription termination is an important regulatory event for both non-coding and coding transcripts. Using high-throughput sequencing, we have mapped RNA Polymerase II's position in the genome after depletion of termination factors from the nucleus. We found that depletion of Ysh1 and Sen1 cause build up of polymerase directly downstream of coding and non-coding genes, respectively. Depletion of Nrd1 causes an increase in polymerase that is distributed up to 1,000 bases downstream of non-coding genes. The depletion of Nrd1 helped us to identify more than 250 unique termination regions for non-coding RNAs. Within this set of newly identified non-coding termination regions, we are further able to classify them based on sequence motif similarities suggesting a functional role for different terminator motifs. The role of these factors in transcriptional termination of coding and/or non-coding transcripts can be inferred from the effect of polymerase's position downstream of given termination sites. This method of depletion and sequencing can be used to further elucidate other factors whose importance to transcription has yet to be determined.

INTRODUCTION

Termination of RNA Polymerase II (Pol II) transcription plays an essential role in the transcription cycle and has been the subject of several recent reviews (Kuehner et al., 2011; Mischo and Proudfoot, 2013). Disruption of the elongation complex at terminators recycles Pol II maintaining a pool of free enzyme able to compete for unoccupied promoters. Correct termination also prevents Pol II from interfering with expression of downstream genes either by colliding with oncoming Pol II elongation complexes or by dislodging transcription factors from the downstream DNA template (Castelnuovo et al.; Cloutier et al., 2013; Shearwin et al., 2005). Termination can also serve a regulatory purpose. Several yeast genes are regulated by premature termination (Arigo et al., 2006a; Creamer et al., 2011; Kuehner and Brow, 2008) and genes involved in yeast nucleotide metabolism are regulated by the choice of alternative transcription start sites, one of which leads to premature termination (Kuehner and Brow, 2008; Thiebaut et al., 2008). More recent studies have shown that correct termination is also necessary for efficient re-initiation at the same gene through the formation of a loop between the 3' and 5' ends (Grzechnik et al., 2014; Tan-Wong et al., 2012).

In addition to mRNAs, Pol II transcribes a diverse set of non-coding RNAs including snoRNA, some snRNAs and several classes of ncRNA with unknown functions. In yeast, this pervasive transcription (Berretta and Morillon, 2009; Jacquier, 2009; Jensen et al., 2013) falls into several classes. Cryptic unstable transcripts (CUTs) are turned over rapidly by the nuclear exosome (Arigo et al., 2006b; Davis and Ares, 2006; Gudipati et al., 2012; Neil et al., 2009; Thiebaut et al., 2006; Wyers et al., 2005). In contrast, another

class of yeast ncRNA, termed stable uncharacterized transcripts (SUTs) are observed in the presence of an active nuclear exosome (Xu et al., 2009).

Pol II terminates coding and non-coding transcripts by different mechanisms (Buratowski, 2005; Hsin and Manley, 2012; Kim et al., 2006; Kuehner et al., 2011; Mischo and Proudfoot, 2013). Coding transcripts and possibly some SUTs are processed at the 3'-end by the cleavage and polyadenylation (pA) machinery. This reaction is coupled to termination occurring downstream of the processing site (Kuehner et al., 2011; Mischo and Proudfoot, 2013). In contrast, ncRNAs are terminated and processed by an alternative pathway that, in yeast, requires the RNA-binding proteins Nrd1 and Nab3 and the RNA helicase Sen1 (Conrad et al., 2000; Steinmetz and Brow, 1996, 1998; Steinmetz et al., 2001; Ursic et al., 2004).

Yeast Pol II terminators contain short RNA sequences that bind proteins within large complexes associated with the elongating Pol II. Loosely conserved pA signal sequences downstream of protein-coding genes bind to components of the CF1 complex leading to assembly of the cleavage and polyadenylation machinery (Kuehner et al., 2011). Termination is coupled to cleavage in a manner that has not yet been completely resolved. Non-pA termination components Nrd1 and Nab3 recognize RNA sequence elements downstream of snoRNAs and CUTs (Carroll et al., 2007; Carroll et al., 2004; Creamer et al., 2011; Porrua et al., 2012; Steinmetz et al., 2001; Steinmetz et al., 2006a; Wlotzka et al., 2011) and this leads to the association of a complex that contains the DNA/RNA helicase Sen1 and the nuclear exosome (Vasiljeva and Buratowski, 2006). The mechanism of termination of these ncRNA transcripts has also not yet been determined.

Several possible mechanisms for Pol II termination have been proposed. The "torpedo" model postulates that cleavage at the pA site exposes an uncapped 5' end on the nascent transcript that acts as a substrate for the 5' → 3' RNA exonuclease Rat1 in yeast or Xrn2 in metazoans (Brannan and Bentley, 2012; Kim et al., 2004b; West et al., 2004). The exonuclease degrades the nascent transcript and upon reaching the Pol II elongation complex facilitates termination by an unknown mechanism. Another model postulates that an allosteric change in Pol II occurs upon assembly of the pA complex (Logan et al., 1987). A member of this complex, Pcf11 has been shown to dismantle an elongation complex in vitro (Zhang and Gilmour, 2006; Zhang et al., 2007) and it is possible that Nrd1 plays a similar role. The DNA/RNA helicase Sen1 interacts with the Pol II CTD (Chinchilla et al., 2012) and has been proposed to act like the bacterial termination factor rho and track along the nascent transcript and pry off the elongating Pol II (Brow, 2011; Kim et al., 1999; Kuehner et al., 2011; Steinmetz and Brow, 1996; Ursic et al., 2004). Mutation of Sen1 has been shown to lead to readthrough of both coding and non-coding transcripts in vivo (Steinmetz et al., 2006b) and in vitro can arrest transcription (Porrua and Libri, 2013).

Part of the uncertainty in delineating termination mechanisms is identifying which factors operate at which terminators in vivo. In this study we have used PAR-CLIP to map Pol II on the yeast genome in living cells (Creamer et al., 2011; Hafner et al., 2010a, b; Jamonnak et al., 2011) after depletion from the nucleus of components of the different termination pathways (Haruki et al., 2008). By comparing the Pol II maps with and without nuclear depletion we are able to map the location of termination to narrow regions of the genome downstream of coding and non-coding genes. This approach

avoids the complexity of using transcripts to map terminators as the effect of RNA turnover is eliminated. We show here that depletion of Ysh1, the protein that cleaves nascent pre-mRNA transcripts at the pA site, enhances accumulation of Pol II at the 3'-ends of protein-coding genes. In contrast, both Nrd1 and Sen1 depletion lead to readthrough transcription of ncRNAs and our data has allowed us to map the position of non-pA termination facilitated by these factors.

RESULTS

Mapping in vivo Pol II termination

To map in vivo termination we have used the anchor-away (AA) system (Haruki et al., 2008) to deplete termination factors from the nucleus. In this approach rapamycin (rap) induces a complex between a protein tagged with a FKBP12-rapamycin binding (FRB) domain and an anchor protein tagged with an FKBP12 domain. In our case the FKBP12 domain is on the ribosomal protein *RPL13A* leading to rap-dependent depletion of FRB-tagged protein from the nucleus. Previous work has shown that a Pol II subunit *RPB1-FRB* strain shows greater than 90% depletion of Pol II at the *PMA1* gene within 40 min (Fan et al., 2010). We show here that in *NRD1-FRBGFP or YSH1-FRBGFP* strains that also contain*, RPB3-TAGRFP,* Nrd1 and Ysh1 are depleted from the nucleus with similar kinetics (Figure 1A-B). We were unable to observe Sen1-FRBGFP because it is present only in ~100-2,000 copies per cell. When over-expressed 10-100-fold from a Gal promoter Sen1-FRBGFP is visible and is depleted from the nucleus with similar kinetics (Figure 1C). Given the limitations of background signals in live-cell imaging we cannot rule out the possibility that some small amount of Nrd1, Ysh1 or Sen1 remain in the nucleus. Growth curves show that all three strains grow normally for several divisions after administration of rap and we carried out PAR-CLIP analysis well before any changes in growth were observed (Figure 2A, shaded box). When cells are plated on media containing rap we observe no growth for *YSH1-FRB* and *NRD1-FRB*, while *SEN1-FRB* grows very slowly (Figure 1D).

Growing yeast cultures were cross-linked as previously described (Creamer et al., 2011; Jamonnak et al., 2011) with modifications described in Methods. Briefly, 4-

thiouracil (4tU) was added to a growing culture for 15 minutes to allow equilibration of uracil pools before the addition of rap. After thirty minutes of rap treatment cultures were irradiated with 365 nm UV for 15 min (Figure 2B). Incubation of yeast cells in 4tU does not significantly affect growth during the course of the experiment as cells continue to grow beyond the time frame of the cross-linking for at least one more doubling (0.2 OD_{600} to 0.7 OD_{600} as measured by BioTek scanner) indicating that 4tU has a less drastic effect on yeast growth than it does in mammalian cells treated for a longer time (Burger et al., 2013).

To map the position of elongating Pol II we isolated Pol II-bound RNA using a dual 6xHis-biotin tagged Pol II Rpb2 subunit (Rpb2-HTB) as previously described (Creamer et al., 2011). Duplicate libraries were derived from the *RPB2-HTB NRD1-FRB* strain grown in the presence or absence of rap. We also created libraries of Rpb2-bound RNA from *SEN1-FRB* and the parental *RPB2-HTB* strains grown in the presence of rap and a single library from *YSH1-FRB*. Replicate libraries were multiplexed and sequenced by Illumina Hi-seq. Each library yielded between 15-39 million unique reads (Table 1). Biological replicates were strongly correlated ($r > .99$) and thus were added together for analysis (Figure 3). The WT dataset correlates very well to NET-seq (Churchman and Weissman, 2011) and GRO-seq (Pelechano et al., 2010) datasets (Figure 4). Thus, PAR-CLIP data accurately represents the position of Pol II elongation complexes on the yeast genome.

Ysh1-dependent termination

To map pA-dependent termination we carried out PAR-CLIP on Rpb2 after depletion of Ysh1, the factor that has been shown to be required for cleavage of the nascent mRNA at the *GAL7* and *CYC1* pA sites (Zhelkovsky et al., 2006). We reasoned that failure to cleave the nascent transcript would prevent Rat1 degradation thus disabling the "torpedo" mechanism. While we cannot rule out the possibility that other factors are depleted along with Ysh1 we do observe a failure to cleave at the pA site for several genes (Figure 5). The most dramatic effect of Ysh1 depletion is seen in a buildup of reads just downstream of the major pA site of highly expressed protein-coding genes (Figure 6A and 2C). The peak of reads is present in most of the highly expressed protein-coding genes as seen in figure 6B. No similar effect is seen downstream of snoRNAs or CUTS. One surprising aspect of Ysh1 depletion is that readthrough Pol II does not extend further downstream but seems to pause or terminate within 200 bp of the pA site (Figure 6A). This result indicates that Pol II that reads through termination signals fails to elongate but is unable to efficiently terminate.

We do, however, observe several instances of Pol II extending 1 kb or more downstream of apparent terminators in response to Ysh1 depletion. The *RNA14* and *DBP2* genes show an increase of Pol II toward the 3' end of the coding region indicating the presence of a Ysh1-dependent terminator upstream of the main pA site (Figure 6D). Rna14p is part of the CF1 complex required for recognition of pA sites (Kuehner et al., 2011) and previous studies have shown that in addition to the mature 2.2 kb mRNA several smaller mRNAs are present under normal conditions (Brendolise et al., 2002; Mandart, 1998). The terminator we map in *RNA14* is located in the same region as the 3' end of the shorter (1 kb) mRNA (Figure 6D).

Dbp2p is a DEAD-box RNA helicase that plays a role in assembly of mRNP complexes and in RNA quality control (Cloutier et al., 2012; Ma et al., 2013). The *DBP2* gene is unique in yeast in having a long (1kb) intron localized toward the 3' end of the coding region and previous work has shown that the gene is autoregulated through sequences in the intron (Barta and Iggo, 1995). In Figure 6D we show that depletion of Ysh1 leads to an increase in Pol II cross-linking downstream of the intron suggesting the presence of an upstream terminator. Consistent with this view polyadenylation sites have been mapped to this location in the Dbp2 intron (Moqtaderi et al., 2013; Ozsolak et al., 2010; Wilkening et al., 2013).

Nrd1-dependent termination

Previous work has shown that Nrd1p is required for proper termination of ncRNAs like snoRNAs and CUTs as well as premature termination of coding transcripts of genes regulated by attenuation (Arigo et al., 2006a; Kuehner and Brow, 2008; Steinmetz et al., 2001; Thiebaut et al., 2008; Thiebaut et al., 2006; Vasiljeva and Buratowski, 2006; Vasiljeva et al., 2008). Here we show that depletion of Nrd1 in the presence of rap leads to readthrough transcription of a large number of these ncRNAs. Similar results have recently been obtained by a different protocol (Schulz et al., 2013).

To map the position of termination we used an approach similar to that used to calculate a travelling ratio (Rahl et al., 2010) or Escape Index (Brannan and Bentley, 2012; Schulz et al., 2013). Reads in 500 bp windows upstream and downstream of a fixed point were tallied and a ratio of reads in the downstream window to total reads in both windows were calculated for that point. This ratio was determined for each point in

the genome on both strands for both control and rap treated cells. Subtracting the fraction of readthrough of the control data from the rap-treated data resulted in a Readthrough Index that allowed us to rank order the regions of the genome showing the highest level of Nrd1-dependent readthrough (Table 2).

Once these regions were identified, we created plots of the difference between the reads in the control and reads with rap (Figure 7A). Nrd1 depletion leads to both a reduction of Pol II upstream and an increase of Poll II downstream of apparent termination sites. This is consistent with previous reports that Pol II pauses prior to termination (Gromak et al., 2006; Hyman and Moore, 1993; Kazerouninia et al., 2010; Larson et al., 2011). In the absence of Nrd1, this apparent pause is eliminated and Pol II is able to transcribe through the termination site. Fitting this difference plot with a spline function allowed us to determine the point at which the difference plot crosses the X-axis. We have called this the termination site but this represents the center point of a narrow (about 50 nt) region over which Pol II is released from the template.

In Figure 7A we show the difference plots for snoRNA *snR34* and the antisense CUT to the *LEO1* gene demonstrating how the termination sites were identified. We have carried out this analysis for 49 snoRNAs and for 144 CUTs showing the highest level of readthrough transcription (Tables S2 and S3). Comparing these termination sites to the sites determined by Pol II ChIP (Schulz et al., 2013) reveals that 70% of the Schulz et al. sites are more than 100 nt downstream of our termination sites and 25% are more than 200 nt downstream (Figure 8). We do not know the reason for this systematic discrepancy but the PAR-CLIP analysis has greater resolution than Pol II ChIP and unlike ChIP gives strand-specificity. The metagene analysis of the ncRNAs for which we

have mapped termination sites is shown in Figure 7B demonstrating a more significant increase in readthrough transcription for Nrd1-depletion when compared to the increase in readthrough transcription due to Ysh1 depletion (Figure 6A). SnoRNA and CUT Nrd1-dependent readthrough extends over 500bp and 1 kb, respectively.

Examining DNA sequences upstream of the termination site for these sets of snoRNAs and CUTs revealed an interesting difference in the occurrence of Nrd1 and Nab3 binding motifs. The top hit in the MEME analysis (Bailey et al., 2009) of the top 27 snoRNA upstream regions (Figure 7C) include the GUA[A/G] sequence previously identified as a Nrd1 binding site (Bacikova et al., 2014; Carroll et al., 2007; Carroll et al., 2004; Creamer et al., 2011; Porrua et al., 2012; Wlotzka et al., 2011). This analysis also identifies a loosely conserved sequence AACUA centered about seven nucleotides upstream of the GUA[A/G] sequence that has not previously been reported. The second most significant motif upstream of the snoRNA termination site was the UCUU sequence that binds Nab3, consistent with the presence of a Nrd1-Nab3 heterodimer (Carroll et al., 2007; Conrad et al., 2000). In sequences upstream of CUT terminators the most significant hit was UCUUG which contains the previously identified Nab3-binding sequence but with a downstream G as has been observed previously (Creamer et al., 2011; Porrua et al., 2012; Wlotzka et al., 2011). Among the 144 CUT terminators the Nrd1 binding motif was not present above background. Together, these observations suggest a significant difference in the manner of recognition of snoRNA and CUT terminators.

U-rich sequences in the Nrd1-dependent termination region

In Figure 9 we show NET-seq (Churchman and Weissman, 2011) and PAR-CLIP data demonstrating that Pol II levels decline in the *SNR13-TRS31* intergenic region. NET-seq reads for several peaks, indicated by bars, are reduced several-fold immediately downstream from the 3'-end of the *SNR13* gene and more than 6-fold further than 100 nt downstream. PAR-CLIP reads from cells grown in the absence of rap decline at the same region reaching a minimum at about 150 nt downstream of the *SNR13* 3'-end. A similar correspondence between NET-seq and PAR-CLIP data is seen for other snoRNA genes (Figure 10). Taken together, these observations indicate that Pol II normally terminates in a Nrd1-dependent fashion in a narrow region located downstream of the snoRNA gene. The sequence from this region for *SNR13*, located from 50 – 150 nt downstream, is shown below the figure and a series of runs of U residues is indicated in red. This is not unexpected as intergenic regions in *S. cerevisiae* are AT-rich. Sequences of termination regions of several antisense CUTs are shown below the *SNR13* sequence. In these cases we also observe runs of U residues. This is unexpected as these termination regions fall within coding regions (on the opposite strand) which in general are not AT-rich. We propose that these U-rich regions constitute part of the non-poly(A) terminator.

Nrd1 and RNA turnover

We estimate that *SNR13* readthrough transcripts are present at less than one copy per cell in wild-type cells and this is consistent with the low level of readthrough transcripts seen in the absence of rap in the Northern blot shown in Figure 11A. Despite this low level of steady-state readthrough transcripts we observe significant cross-linking to Rpb2 in the region downstream of *SNR13* (Figure 11B) even in the presence of Nrd1. Presumably the RNA synthesized by these polymerases is rapidly degraded by the

nuclear exosome. Figure 11A shows that snR13 readthrough transcripts detected by Northern blot or RT-PCR are increased more than 10-fold by 15 min and more than 100-fold by 30 min of Nrd1 depletion. However, depletion of Nrd1 results in only a 2-3-fold increase in PAR-CLIP signal for Rpb2 (Figure 11B) indicating that depletion of Nrd1 not only allows readthrough of ncRNA transcripts but also results in their stabilization by uncoupling degradation by the nuclear exosome.

Our analysis of steady-state RNA levels is consistent with recently published 4tU-seq data (Schulz et al., 2013). Figure 11C shows the comparison of our data and that of Schulz et al for *SNR13* and the upstream *RPL27B* gene and for the *SNR62* and *SNR9* genes. 4tU-seq data show a dramatic increase in *SNR13* and *SNR62* readthrough transcription compared to PAR-CLIP data. The most likely explanation for this difference is that 4tU labeling is not completely restricted to nascent transcripts. Labeling cells for six minutes with 4tU allows for multiple rounds of transcription especially for heavily transcribed genes like snoRNAs and ribosomal protein genes. The earliest synthesized transcripts are subject to processing as is clearly evident in the paucity of reads derived from the intron of *RPL27B* in the 4tU-seq compared to the PAR-CLIP (Figure 11C) and a number of other intron-containing genes (Figure 12). Depletion of Nrd1 does not effect all snoRNAs as is seen in Figure 11C where neither PAR-CLIP nor 4tU-seq detect appreciable readthrough transcription at *SNR9*.

Sen1-dependent termination

A role for Sen1p in termination of both coding and ncRNA transcripts has been proposed (Steinmetz and Brow, 1996; Steinmetz et al., 2001; Steinmetz et al., 2006b). Our current

data show that depletion of Sen1 results in no change in the localization of Pol II at the 3'-ends of most protein-coding genes (Figure 13A) including those previously shown (Steinmetz et al., 2006b) to be dependent on Sen1 (Figure 14). Thus, despite our previous observation that Sen1 cross-links to these transcripts (Creamer et al., 2011), there is little evidence for a role for Sen1 in pA termination. Sen1 depletion does result in readthrough transcription at both snoRNAs and CUTS (Figure 13B-C) which is consistent with previous observations (Arigo et al., 2006b; Steinmetz et al., 2001; Thiebaut et al., 2006). The pattern of Pol II readthrough in response to Sen1 depletion differs, however, from that of Nrd1. Rather than an extended region of readthrough downstream of the terminator, Sen1 depletion often results in an increase just downstream by a few hundred nucleotides. This type of pattern is seen in Figure 13D for *SNR47* and the CUT antisense to the *COG8* gene and in Figure 15 for several protein-coding genes regulated by Nrd1-dependent attenuation. Figure 13E compares the percent readthrough downstream of all 266 non-pA terminators in Sen1 depleted and Nrd1 depleted cells. The percent readthrough after Nrd1 depletion is twice that for Sen1 depletion and these values are not dominated by a subset of terminators. This observation suggests that Sen1 may play a different function than Nrd1 at these genes and that loss of Sen1 results in less processive or more pause prone readthrough transcription.

DISCUSSION

In this paper we describe the development and implementation of a technique for mapping the position of in vivo Pol II termination. Depletion of components of the yeast Pol II termination machinery followed by Pol II cross-linking to nascent RNA has allowed the identification of both pA and non-pA termination regions. The site of pA termination has been mapped to within a few hundred bases of the pA site as has previously been observed for the Cyc1 terminator (Birse et al., 1998). We have identified non-pA terminators using an algorithm that identifies points in the genome where Pol II readthrough increases in response to Nrd1 depletion.

While many of the components of both the pA and non-pA termination complexes have been identified, the mechanism of termination is still unclear. Two general models have been proposed (Buratowski, 2005; Kuehner et al., 2011; Luo and Bentley, 2004; Mischo and Proudfoot, 2013). The allosteric model proposes a conformational change in Pol II that slows elongation (Calvo and Manley, 2001; Logan et al., 1987; Nag et al., 2006) while the torpedo model proposes that cleavage at the pA site creates a substrate for a 5' → 3' exonuclease (Rat1 in yeast and Xrn2 in metazoa) that degrades the nascent transcript like a "fuse" and upon reaching elongating Pol II somehow facilitate termination (Connelly and Manley, 1988; Kim et al., 2004b; West et al., 2004). Neither model adequately explains all experimental observations leading to a hybrid model that includes both allosteric and topedo mechanisms (Luo et al., 2006).

pA Termination

Our data on Ysh1 depletion supports a two-step pause and release model for termination that invokes both allosteric and torpedo mechanisms (Nag et al., 2007). The buildup of

Pol II just downstream of the uncleaved pA site suggests that Pol II has paused but is deficient in the subsequent release step. This is consistent with a role for Ysh1 in providing the substrate for the Rat1 exonuclease that facilitates removal of Pol II from the template (Pearson and Moore, 2013). Ysh1 apparently is not required for the allosteric change that results in pausing downstream of the pA site as seen in the lack of readthrough at most pA sites. In contrast, mutation of the pA site (Kim et al., 2004a) or mutations in *RNA14*, *RNA15*, *PCF11*, *CLP1* and *GLC7* result in readthrough transcription more than 500 nt downstream (Al Husini et al., 2013; Birse et al., 1998; Schreieck et al., 2014) arguing for a role for these factors in the allosteric change that renders Pol II immobilized.

Depletion of Ysh1 has also revealed several genes that are apparently regulated by premature termination through the pA pathway. *RNA14* and *DBP2* contain pA sites upstream from the 3' pA site. Loss of Ysh1 leads to processive elongation ending at the downstream pA site indicating that Ysh1 is required for termination at these upstream sites. Why the putative allosteric step does not function efficiently at these upstream sites is not clear. Perhaps the phosphorylation pattern of the CTD at these sites precludes the assembly of factors like Pcf11. Alternatively, chromatin structure modification induced by the Pol II that elongates through these terminators may suppress normal termination. Finally, it is possible that use of the downstream terminator by Pol II that ignores the upstream pA site creates a gene loop that favors termination at the downstream site.

Non-pA Termination

Nrd1 and Nab3 bind specific RNA sequences and act as sensors to detect non-pA terminator sequences in the nascent transcript. While a number of studies have characterized the short motifs recognized by Nrd1 and Nab3 the relative orientation and abundance of these sequences varies widely among non-pA terminators (Carroll et al., 2007; Carroll et al., 2004; Creamer et al., 2011; Jamonnak et al., 2011; Porrua et al., 2012; Schulz et al., 2013; Wlotzka et al., 2011). In this study we have localized termination downstream of 49 snoRNAs and 144 CUT transcripts. This has allowed a search for sequences that may define these two sets of terminators. We find that the most significant motif in this set of snoRNA terminators contains the GUA[A/G] motif previously identified for Nrd1 in the context of a longer sequence that indicates the possible involvement of another, unidentified RNA-binding protein. The second most significant motif is the Nab3 binding sequence UCUUG. In the case of the CUT terminators the only significant motif is the Nab3 binding sequence. Thus, it appears that while both Nrd1 and Nab3 binding contribute to recognition of snoRNA terminators Nab3 predominates for CUTs. This does not preclude Nrd1 from playing a critical role in termination at these CUTs. We have previously shown that Nrd1 is necessary for termination of the CUT antisense to *FMP40* presumably by acting as an adaptor to couple the Nrd1-Nab3-Sen1 complex to the elongating Pol II through interaction between the Nrd1 CID and the CTD (Arigo et al., 2006b).

No other significant motifs were observed upstream of these terminators but we did observe that the sequences surrounding the termination site are U-rich and contain multiple runs of U residues. This is not unexpected for snoRNA downstream sequences as intergenic regions are AT-rich in *S. cerevisiae*. The U-rich sequences surrounding the

termination sites (Figure 9) of antisense CUTs are more significant as these sequences occur in the context of a coding region on the opposite strand. U-rich sequences have been shown to form unstable rU:dA base pairs (Huang et al., 2010) and in the hybrid binding site of Pol II transcribing the antisense CUT such an unstable hybrid sequence may increase the probability of termination (Kireeva et al., 2000; Martin and Tinoco, 1980).

In addition to its role in Pol II termination, our data highlight the role Nrd1 plays in the turnover of ncRNAs. Nrd1 depletion by anchor away leads to accumulation of readthrough transcripts in 4tU-seq to a much higher degree than observed by our Rpb2-HTB PAR-CLIP. We show here by Northern and RT-PCR that after 40 min of Nrd1 depletion the level of snR13 readthrough transcripts increases over 100-fold. However, after 40 minutes of depletion the amount of readthrough as observed by PAR-CLIP is less than 10-fold. This difference is likely due to an important role for Nrd1 in coupling termination to turnover of the completed transcript. Nrd1 is found in a complex with components of the TRAMP and exosome complexes (Vasiljeva and Buratowski, 2006) and previous studies have indicated that *nrd1* mutants are deficient in RNA turnover (Arigo et al., 2006b; Thiebaut et al., 2006; Vasiljeva and Buratowski, 2006).

Pulse labeling RNA with 4tU was recently used to analyze newly synthesized transcripts in a Nrd1 anchor away mutant and this data identified over 1500 Nrd1-dependent transcripts (Schulz et al., 2013), far more than the number of terminators we describe here. Part of the reason for this discrepancy is that the six-minute pulse used in the 4tU-seq protocol is substantially longer than the time needed to synthesize short

RNAs. Thus, 4tU-seq over-estimates the effect on termination and may ascribe termination functions to cases where the role in turnover is independent of termination.

Depletion of the DNA/RNA helicase Sen1 has allowed us to place this factor squarely in the Nrd1-Nab3 non-pA termination pathway. We observe readthrough transcription downstream of both snoRNAs and CUTs but not downstream of pA sites. Pol II does not progress far downstream after Sen1 depletion indicating that this factor may act after the allosteric change has occurred. This is consistent with Sen1 acting in place of Rat1 to provide the activity that dislodges the paused Pol II (Porrua and Libri, 2013) and supports previous work showing the difference in the requirement for these factors in pA and non-pA termination (Kim et al., 2006).

The results presented here are summarized in a model shown in Figure 16. We have been able to demonstrate a difference in the global effect of depleting different components of the yeast termination machinery. These results have been interpreted in the context of a two state model for termination in which terminator sequences in the nascent RNA trigger assembly of protein complexes that signal to the elongating Pol II to decrease its elongation rate and processivity. These anti-processivity factors include Pcf11 for pA terminators and Nrd1 for non-pA terminators. Loss of either of these factors leads to runaway processive elongation. In contrast, depletion of Ysh1 or Sen1 does not block the transition to non-processive elongation but does result in a defect in removal of Pol II from the template. For Sen1 this likely occurs through loss of a rho-like function (Brow, 2011; Porrua and Libri, 2013) while Ysh1 depletion leads to a lack of cleavage at the pA site thereby limiting access of Rat1 and preventing the torpedo mechanism of Pol II release (Pearson and Moore, 2013).

The data presented here demonstrate that PAR-CLIP analysis of Pol II in living yeast cells can define the site of Pol II termination regions and when coupled with depletion of termination factors offers an avenue for examining the roles of different factors in termination genome-wide. Such experiments will illuminate the variety of mechanisms the cell uses to prevent Pol II from transcribing beyond genetically determined 3' boundaries.

MATERIALS and METHODS

Strain Construction

Yeast strains for anchor away were constructed using the parental strain HHY168 (Haruki et al., 2008) (Euroscarf #Y40343). *RPB2* was C-terminally tagged with an HTB (6xHIS, TEV, and Biotin) tag as described previously (Creamer et al., 2011; Tagwerker et al., 2006). *NRD1, YSH1, and SEN1* were C-terminally tagged with FRB and FRBGFP as described in Haruki et al. (Haruki et al., 2008). The SEN1 promotor was replaced with the *GAL1* promotor using a cassette from pFA6a-kanMX6-PGAL1-HBH (Tagwerker et al., 2006). All strains used in these experiments are listed in Table 4.

Cell Growth

Cells were grown in Complete Synthetic Media (CSM) supplemented with 2% glucose or galactose and 40 mg/l adenine. For RNA analysis, cells were seeded to an OD_{600} of 0.1 from a 5 ml or 50 ml overnight culture. The cells were then incubated at 30°C to an OD_{600} of 1. 1 μg/ml of rap (LC laboratories) was then added to the cultures and the cells were allowed to grow for the indicated times in the experiment. For growth curves, a BioTek Infinity 2 (BioTek) was used to incubate a 24 well plate containing CSM media with and without 1 μg/ml rap (added after 6 hrs of growth) with orbital shaking (Slow, 1 mm amplitude). Cells were diluted to an OD_{600} of 0.1 per well and an OD_{600} was measured every 10 min for 24 hours.

RNA Purification and cDNA synthesis

Total RNA was extracted from yeast with hot acid phenol as previously described (Creamer et al., 2011). Strand specific northern blots were done as described previously

(Marquardt et al., 2011) with the following modifications. Ultrahyb (Invitrogen) was used instead of the described hybridization buffer. Real-time PCR analysis was done on a CFX96 instrument (Biorad) in triplicate as described previously (Jamonnak et al., 2011). Primers used can be found in Table 5.

Live Cell Imaging of FRBGFP Strains.

The strains containing *NRD1-FRBGFP, YSH1-FRBGFP, or GALpSEN1-FRBGFP, RPB3-TAGRFP* were grown to an OD_{600} of 1 at 30°C. Live cell imaging was done on a Deltavision microscope (Applied Precision) as described previously (Rines et al., 2011) with the following modification: 1 μg/ml rapamyacin was added to the agarose/media pad. Galactose was substituted for glucose with the GALpSEN1-FRBGFP strain. Pictures were taken of GFP at 0 mins and 45 mins post rap exposure and 45 mins for RFP.

Anchor Away and PAR-CLIP

Cells were grown overnight 30°C in 50 ml of YPD (Yeast Extract Peptone with 2% glucose) media. Four 500 ml flasks of sterile cross-linking media (CSM–Ura supplemented with 2% glucose, 40 mg/l adenine, 60 μM uracil, and 1 μM biotin were seeded to an OD_{600} of 0.1 and incubated at 30°C until an OD_{600} of 1. Crosslinking was carried out by the protocol described in Figure 2B. 4-thiouracil (4tU) was added to a final concentration of 4 mM and incubated at 30°C for 15 min. 1 μg/ml of rapamycin was then added to the cells and incubated at 30°C for 30 min. All four 500 ml cultures were then pooled into one 2 l beaker and irradiated from a distance of two cm with 365 nm UV (~ 1 W/cm^2) from an LC-L5 LED UV Lamp (Hamamatsu) for 15 min with gentle stirring. The top ~2 mm of the culture appears luminescent indicating that the UV light does not

penetrate further into the culture. Assuming an even mixture of the culture we calculate that the average cell is cross-linked for only ~ 10 seconds within the 15 min crosslinking period. The cultures were then filtered through 0.45 micron nitrocellulose filters (Millipore) and cells were scraped into 5 ml Buffer 1 (300 mM NaCl, 0.5% NP-40, 50 mM $NaPO_4$ pH 7.2, 10 mM immidazole, 6 M Guanidine HCl, Protease inhibitor Cocktail VII [RPI]) and frozen in liquid nitrogen.

HTB tagged Rpb2 purification was adapted from our previously published protocol (Jamonnak et al., 2011). Cells were lysed in liquid nitrogen using a SamplePrep 6870 freezer mill (Spex) with 15 cycles per second of cracking for 15 cycles of 1 min with 2 min of cooling between each cycle. Lysates were then incubated with 1 ml of Ni-NTA agarose (Qiagen), which was equilibrated in buffer 1 for 2 hours at room temperature. The Ni-NTA agarose was then added to an empty 10 mL plastic column (Biorad) and washed with 20 ml of Buffer 1 followed by 10 ml of Buffer 2 (20 mM $NaPO_4$ pH 7.2, 300 mM NaCl, 0.5% NP40, 10 mM imidazole, 4 M Urea). The protein was eluted off of the Ni-NTA agarose with 5 ml Buffer 2 + 250 mM immidazole and into 5 ml Buffer 2 + 120 µl Protease inhibitor Cocktail VII + 100 µl of hydrophilic streptavidin magnetic beads (NEB). The slurry was allowed to incubate for 4 hours at room temperature. The strepavadin magnetic beads were then resuspended in 500 ul of Buffer 3 (50 mM Tris pH 7.4, 200 mM NaCl, 4 M Urea), transferred to a 1.5 ml siliconized tube and washed with 3 x 1 ml of Buffer 3 followed by 3 x 1 ml of T1 Buffer (50 mM Tris pH 7.4, 150 mM NaCl, 2 mM EDTA). Beads were resuspended in 200 µl of T1 buffer. 0.15 U/µl of Rnase T1 (Fermentas) was added to the bead slurry and allowed to incubate at 25°C for 15 min. The beads were then washed 3 x 1 ml T1 wash buffer (50

mM Tris pH 7.4, 500 mM NaCl, 1% NP-40, 0.5% Na deoxycholate) followed by 3 x 1 ml washes in PNK Buffer (50 mM Tris pH 7.2, 50 mM NaCl, 10 mM $MgCl_2$). The beads were resuspended in 200 µl of TSAP reaction solution (0.15 U/ul Thermosensitive Alkaline Phosphatase [TSAP, Promega], 1 U/µl SupeRase Inhibitor [Invitrogen], 1 mM DTT in PNK Buffer) and incubated at 37°C for 30 min. The beads were washed 1 x 1 ml Buffer 3, 2 x 1 ml T1 Wash Buffer, and 3 x 1 ml PNK Buffer.

cDNA Library Preparation

Streptavadin beads were resuspended in 200 ul of PNK reaction solution (1 U/µl T4 PNK [NEB], ^{32}P g-ATP, 5 mM DTT in PNK Buffer) and incubated at 37°C for 30 min. 1 mM of cold ATP was added to the reaction and incubated at 37°C for 10 min. The beads were washed 1 x 1 ml Buffer 3, 1 x 1 ml T1 wash, 3 x 1 ml T4 RNA Ligase Buffer (50 mM Tris pH 7.4, 10 mM $MgCl_2$). The beads were resuspended in 44 µl of T4 RNA Ligase 2 reaction solution (25% PEG 8000, 5 µM 3' adaptor [AppAGATCGGAAGAGCACACGTCTddC, IDT], 10 U/µl T4 RNA Ligase 2, truncated K227Q [NEB], 2 U/µl RNase Inhibitor [Invitrogen], 1 mM DTT in RNA Ligase Buffer) and incubated at 25°C for 4 hours. The beads were then washed with 1 x 1 ml Buffer 3 and 5 x 1 ml T4 RNA Ligase Buffer. The beads were resuspended in 50 ul of T4 RNA Ligase reaction solution (5 uM 5' Adaptor [GUUCAGAGUUCUACAGUCCGACGAUC, IDT], 1.2 U/µl RNase Inhibitor, 1mM ATP, 0.5 U/µl T4 RNA Ligase [NEB], 1 mM DTT in RNA Ligase Buffer) and incubated at 16°C overnight. The beads were then washed 5 x 1 ml Proteinase K Buffer (100 mM Tris pH 7.4, 150 mM NaCl, 12.5 mM EDTA). The beads were resuspended in 200 µl of Proteinase K Buffer + 2% SDS and 12 µl of Proteinase K (NEB). The suspension was incubated for 30 min at 37°C at which point the

supernatant was removed and saved and the beads were resuspended in 200 µl of Proteinase K Buffer + 2% SDS and 12 µl of Proteinase K. The beads were then incubated 30 min at 37°C and the supernatants were pooled. The RNA was recovered by acid phenol/chloroform extraction followed by two serial ethanol precipitations. The resulting pellet was allowed to dry and resuspended in 15 µl of Nuclease-Free Water [Invitrogen]. The RNA was split into 5 µL aliquots and frozen at -80°C for storage.

The Reverse Transcription, PCR, and gel extraction of the cDNA Library were carried out as described previously (Jamonnak et al., 2011) with the following modifications. For reverse transcription, the primer used was AGACGTGTGCTCTTCCGATCT (IDT). For PCR analysis, biological repeats were multiplexed with the following primers. The forward primer was AATGATACGGCGACCACCGAGATCTACACGTTGAGAGTTCTACAGTCCG*A (where a * denotes and phosphorothioate bond). The first reverse primer was CAAGCAGAAGACGGCATACGAGATATTGGCGTGACTGGAGTTCAGACGTGTG CTCTTCGGATC*T for the first biological repeat and the second reverse primer was CAAGCAGAAGACGGCATACGAGATTACAAGGTGACTGGAGTTCAGACGTGT GCTCTTCCGATC*T for the second biological repeat (IDT).

Sequencing and Bioinformatics

Sequencing and demultiplexing was done at UC Riverside on an Illumina HiSeq. Trimming of the resulting sequences as described previously (Jamonnak et al., 2011). Briefly, raw reads were trimmed of the 3' adaptor using a wrapper for R-bioconductor (Gentleman et al., 2004) developed by Sarah Wheelan. Raw trimmed reads were

condensed (defined as no more than one of the same exact sequence) to eliminate any PCR artifacts within the data. Bowtie 1.0.0 (Langmead, 2010) was run with the following arguments (-y --best -v 2) and aligned to the SacCer3 (R64) genome. Reads mapping to tRNA genes were removed as these likely represent artifactual binding during the affinity purification steps. Reads were then converted to a wig format where each read was multiplied by a factor that was defined as the total number of reads aligned per 10^7 reads. PAR-CLIP datasets have been submitted to GEO with the accession number GSE56435.

Calculation of the Top 500 PolyA Sites.

To calculate the best polyA sites per gene we used the sites provided by Moqtaderi et al. (Moqtaderi et al., 2013). We set a hard cutoff of at least 200 raw reads. If, however, there were multiple very strong polyA sites within 200 base pairs of each other, we calculated the read ratio between the two most significant polyA sites. If the read ratio was less than 0.4, we ignored them.

Global Readthrough Percentage

Global readthrough percentage was calculated for each point in the genome as the percentage readthrough (reads downstream 500 bp / reads downstream 500 bp + reads upstream 500 bp) in the treated sample minus control sample. All points with a percentage readthrough of greater than 10%, at least a total of 1000 reads in either sample, and greater reads in the treatment vs the control were extracted. The list was further refined by taking the difference of every point 2 kb around non-overlapping termination sites (approx 500 total), graphing the region, and then fitting a spline function over a 25

bp window to find the best point of inflection. These points of inflection were then used as a focal point for the meta gene analysis.

Acknowledgements

We thank Dr. Sarah Wheelan (Johns Hopkins) for the 3' trimming program and Dr. Andrew Holland (Johns Hopkins) for use of the microscope. Professor Patrick Cramer (Gottingen) provided access to 4tU-seq data and helpful comments on the manuscript. Jennifer Fairman (Johns Hopkins) provided illustrations. Dr. Nicholas Guydosh (Johns Hopkins) provided helpful comments on the manuscript.

Figure 1. Nuclear depletion of Pol II termination factors.

Live yeast cell imaging of: **A.** *NRD1-FRBGFP, RPB3-TAGRFP;* **B.** *YSH1-FRBGFP, RPB3-TAGRFP;* **C.** *GALpSEN1-FRBGFP, RPB3-TAGRFP* strains. Time after rap treatment is indicated above the pictures. **D**. Growth of indicated FRB strains on CSM plates with and without rap.

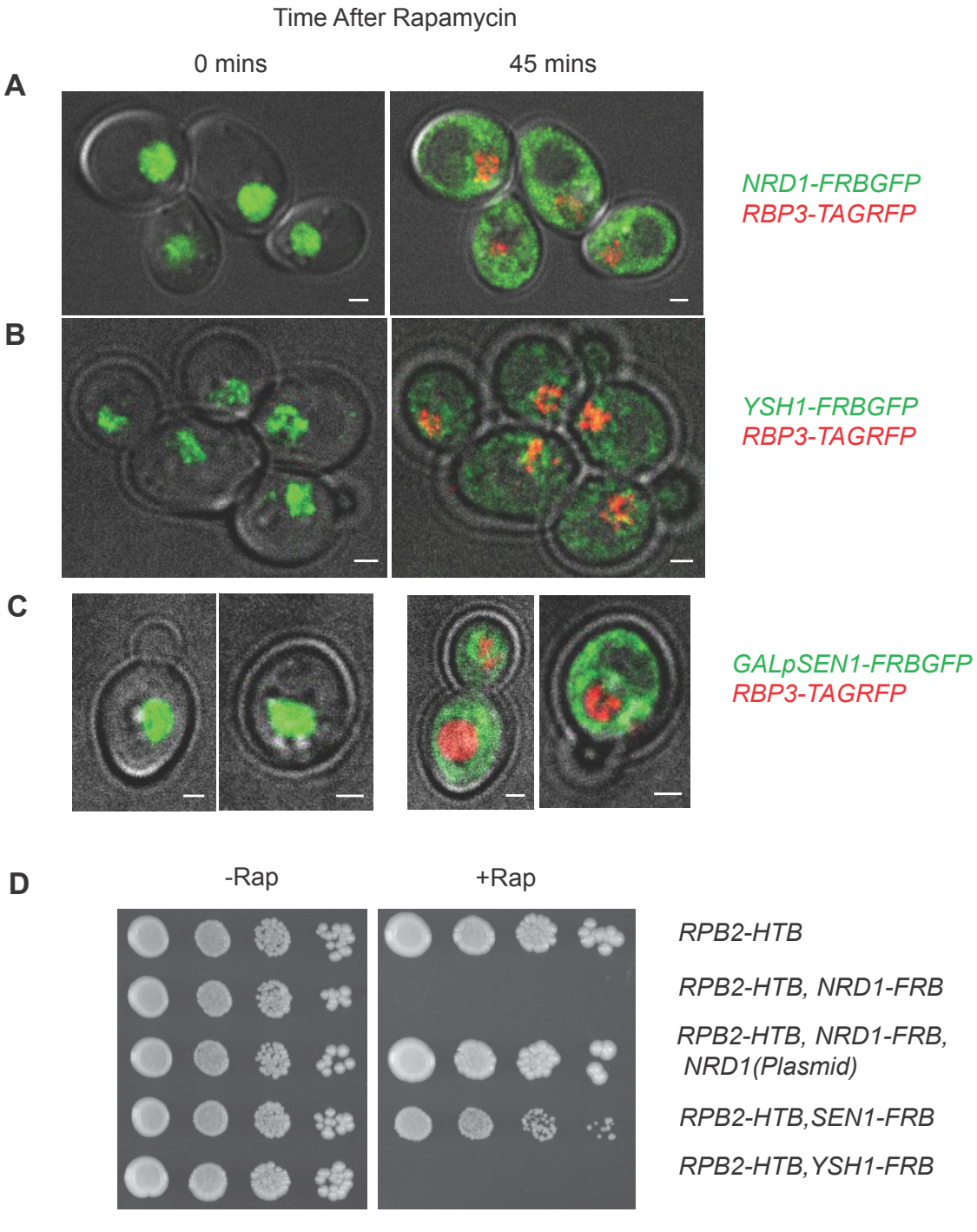
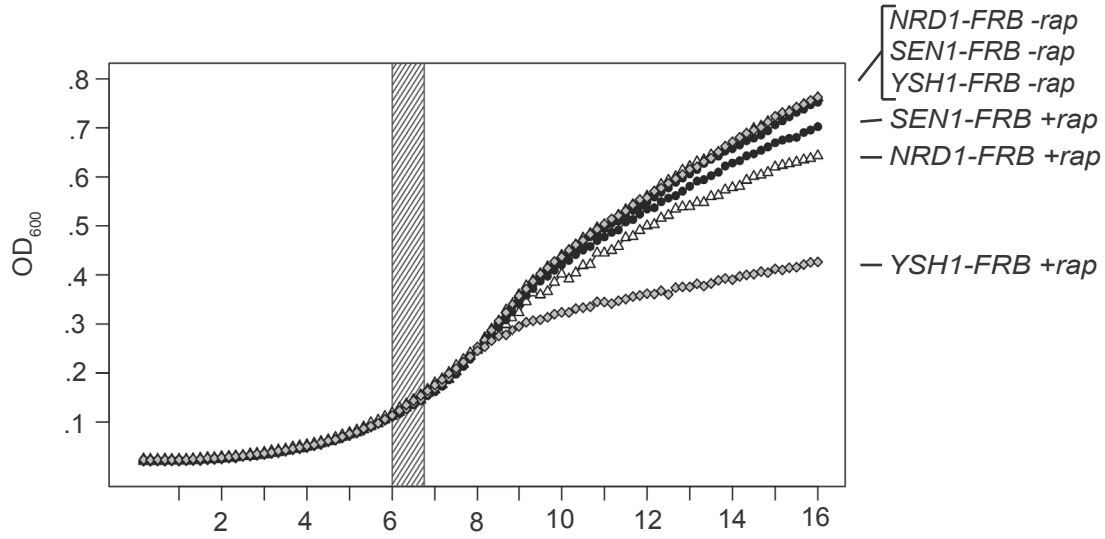

Time After Rapamycin
0 mins
45 mins
A
NRD1-FRBGFP
RBP3-TAGRFP
B
YSH1-FRBGFP
RBP3-TAGRFP
C
GALpSEN1-FRBGFP
RBP3-TAGRFP
D
-Rap
+Rap
RPB2-HTB
RPB2-HTB, NRD1-FRB
RPB2-HTB, NRD1-FRB,
NRD1(Plasmid)
RPB2-HTB,SEN1-FRB
RPB2-HTB,YSH1-FRB

Figure 2. Growth of yeast strains.

A. Growth of indicated FRB strains in CSM media with and without rap. OD_{600} was taken every 10 min for 20 hours. The grey bar represents the timeline of the experiment. **B.** Timeline of the PAR-CLIP protocol.

A


NRD1-FRB -rap
SEN1-FRB -rap
YSH1-FRB -rap
SEN1-FRB +rap
NRD1-FRB +rap
YSH1-FRB +rap
OD600
.8
.7
.6
.5
.4
.3
.2
.1
2
4
6
8
10
12
14
16

B

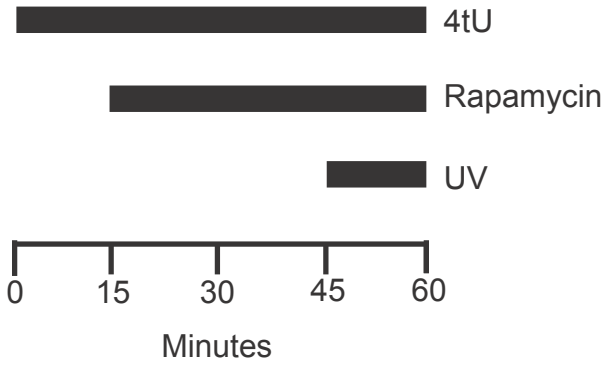

4tU
Rapamycin
UV
0
15
30
45
60
Minutes

Figure 3. Comparison of PAR-CLIP biological replicates

A-D. Replicate pair-wise comparison of number of reads corresponding to each gene in our PAR-CLIP biological replicates. Spearman coefficient is represented by rho.

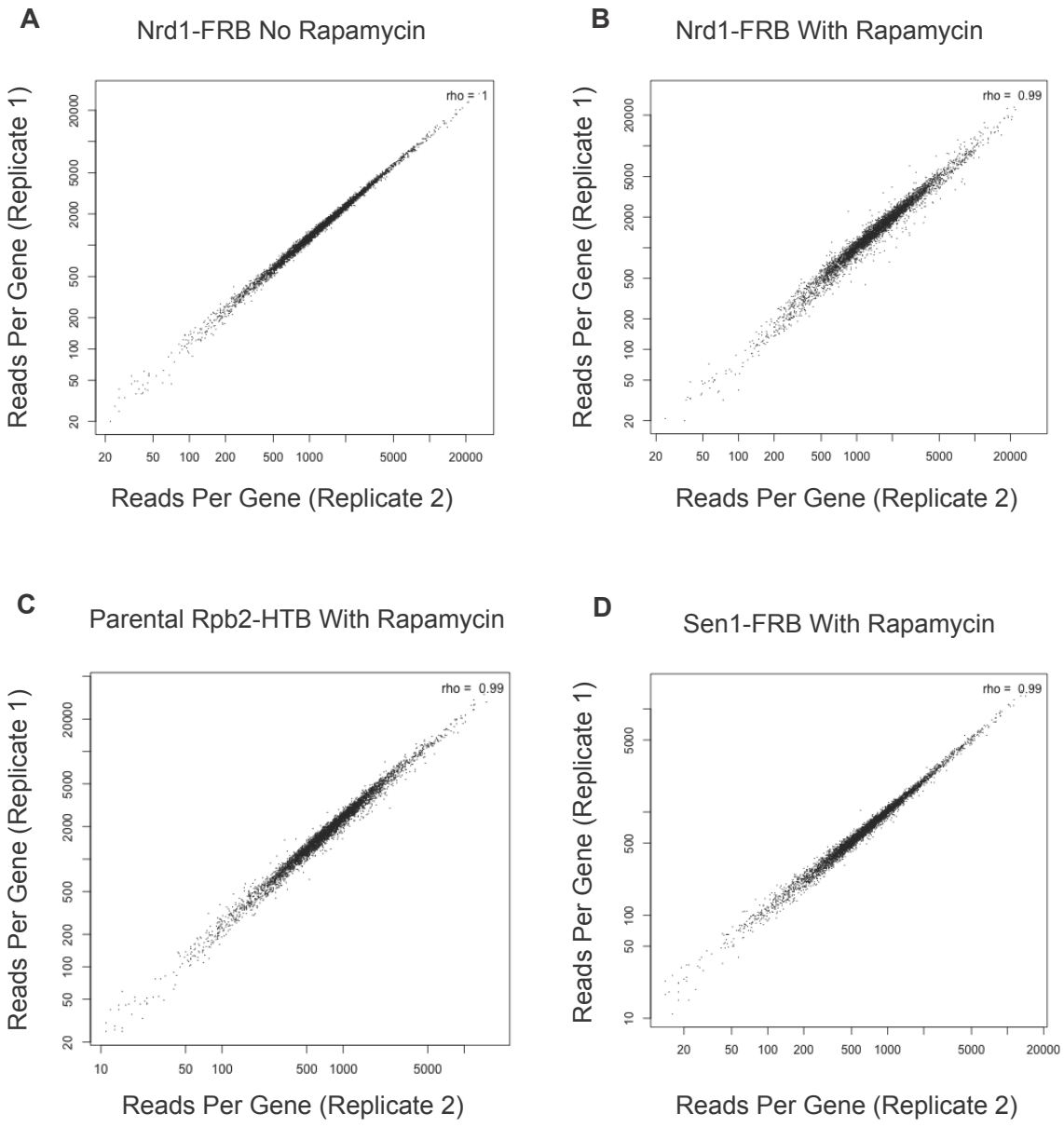

A
Nrd1-FRB No Rapamycin
rho = 1
Reads Per Gene (Replicate 1)
Reads Per Gene (Replicate 2)
B
Nrd1-FRB With Rapamycin
rho = 0.99
Reads Per Gene (Replicate 1)
Reads Per Gene (Replicate 2)
C
Parental Rpb2-HTB With Rapamycin
rho = 0.99
Reads Per Gene (Replicate 1)
Reads Per Gene (Replicate 2)
D
Sen1-FRB With Rapamycin
rho = 0.99
Reads Per Gene (Replicate 1)
Reads Per Gene (Replicate 2)

Figure 4. Comparison of PAR-CLIP with other datasets.

A-C. Pair wise comparison of number of Rpb2-HTP cross-linked reads in each annotated gene vs. GRO-seq reads [54] or NET-seq reads [53]. Spearman coefficient is represented by rho.

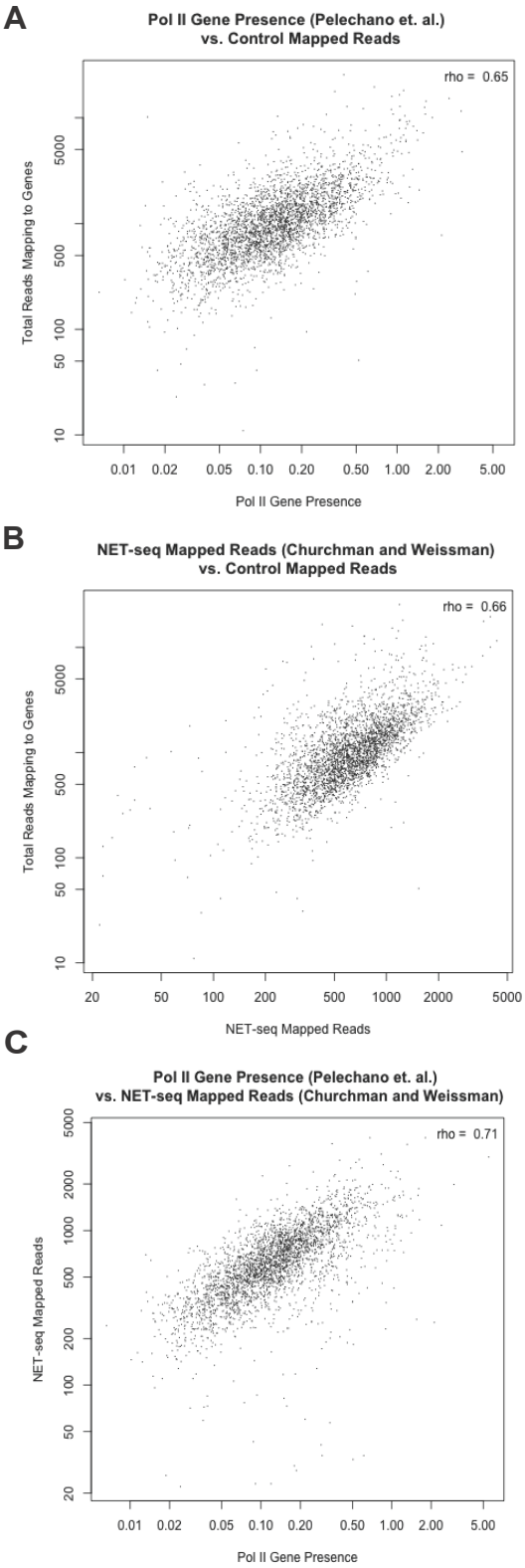

A
Pol II Gene Presence (Pelechano et. al.)
vs. Control Mapped Reads
rho = 0.65
Total Reads Mapping to Genes
10
50
100
500
5000
0.01
0.02
0.05
0.10
0.20
0.50
1.00
2.00
5.00
Pol II Gene Presence
B
NET-seq Mapped Reads (Churchman and Weissman)
vs. Control Mapped Reads
rho = 0.66
Total Reads Mapping to Genes
10
50
100
500
5000
20
50
100
200
500
1000
2000
5000
NET-seq Mapped Reads
C
Pol II Gene Presence (Pelechano et. al.)
vs. NET-seq Mapped Reads (Churchman and Weissman)
rho = 0.71
NET-seq Mapped Reads
20
50
100
200
500
1000
2000
5000
0.01
0.02
0.05
0.10
0.20
0.50
1.00
2.00
5.00
Pol II Gene Presence

Figure 5. Verification of readthrough of Pol II after Ysh1 depletion

A-D. Histograms representing normalized reads with Ysh1 (grey) and without Ysh1 (black) at the given genomic locations. Genes and pA sites are represented below each graph and the length of the genome depicted is given in the lower right hand corner. Below are qPCR at different times after addition of rap using the amplicon highlighted in red.

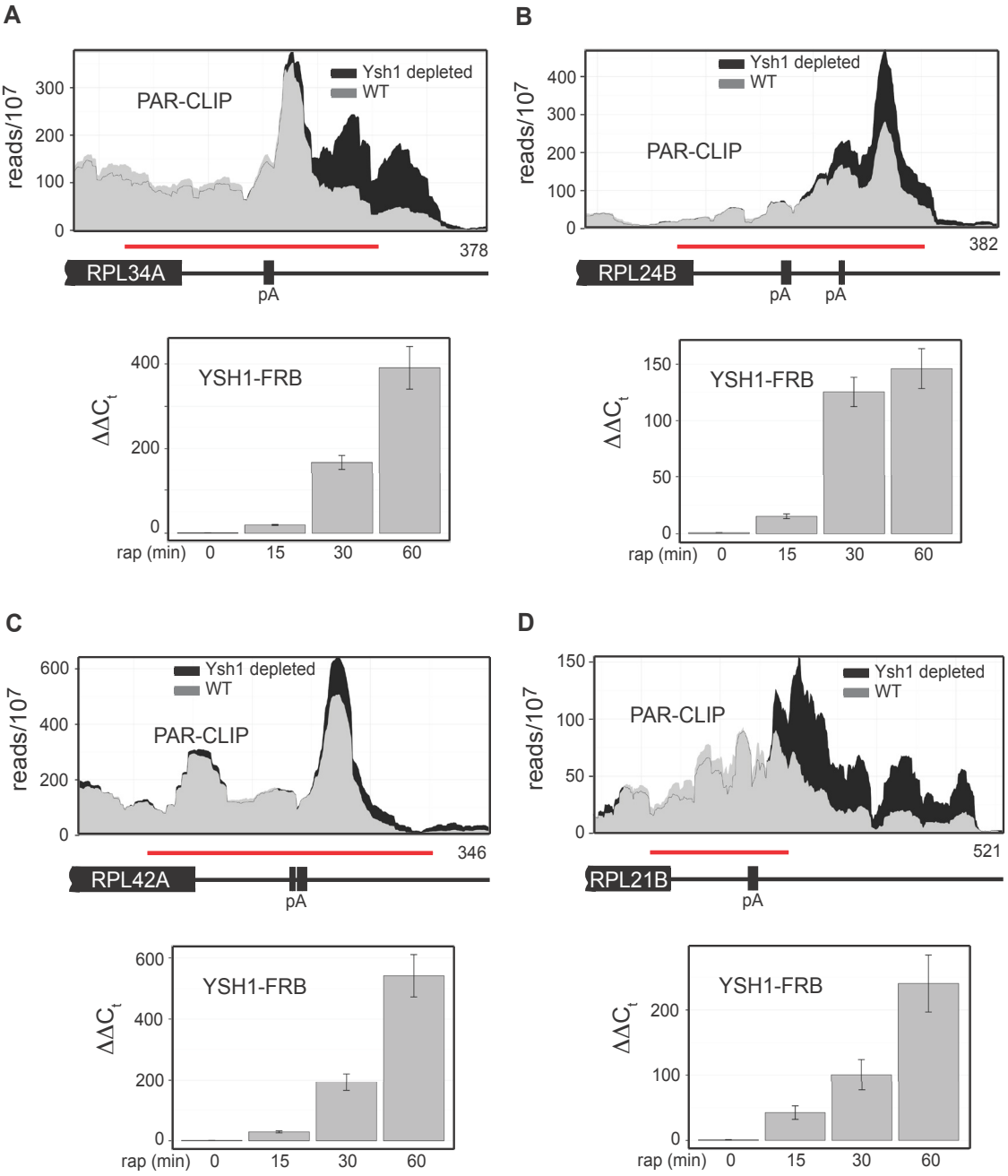

A
reads/10^7
PAR-CLIP
Ysh1 depleted
WT
378
RPL34A
pA
YSH1-FRB
ΔΔC_t
rap (min) 0 15 30 60
B
reads/10^7
PAR-CLIP
Ysh1 depleted
WT
382
RPL24B
pA
pA
YSH1-FRB
ΔΔC_t
rap (min) 0 15 30 60
C
reads/10^7
PAR-CLIP
Ysh1 depleted
WT
346
RPL42A
pA
YSH1-FRB
ΔΔC_t
rap (min) 0 15 30 60
D
reads/10^7
PAR-CLIP
Ysh1 depleted
WT
521
RPL21B
pA
YSH1-FRB
ΔΔC_t
rap (min) 0 15 30 60

Figure 6. Ysh1 depletion causes readthrough at pA sites.

A. Mean reads every 10 bp for the 500 most frequently used pA sites with (dotted) and without (solid) Ysh1. **B**. Plot showing percent readthrough of each of 500 pA terminators as a fraction of the total 500 pA terminators. **C and D**. Histograms representing normalized reads with Ysh1 (grey) and without Ysh1 (black) at the given genomic locations. The TSS with the direction of transcription is indicated by an arrow. Genes and pA sites are represented below each graph and the length of the genome depicted is given in the lower right hand corner.

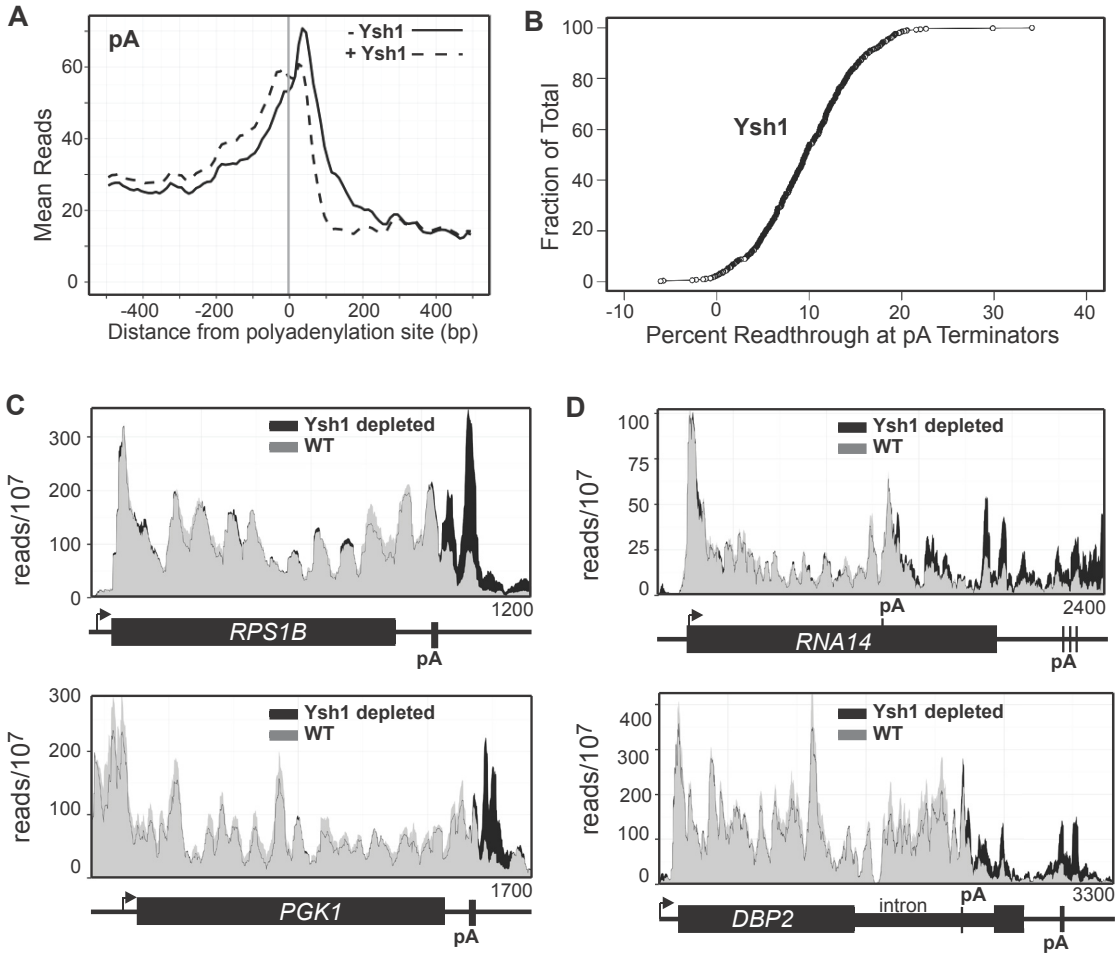

A
pA
- Ysh1
+ Ysh1
Mean Reads
Distance from polyadenylation site (bp)
B
Ysh1
Fraction of Total
Percent Readthrough at pA Terminators
C
Ysh1 depleted
WT
reads/10^7
RPS1B
pA
1200
PGK1
1700
D
RNA14
2400
DBP2
intron
3300

Figure 7. Mapping Nrd1-dependent terminators.

A. Histograms representing of normalized reads for *NRD1-FRB* with Nrd1 (grey) and without Nrd1 (black) at the given genomic locations. The difference between WT and Nrd1-depleted at every nucleotide is represented as a histogram below each graph. **B.** Mean reads every 10 bp for 49 snoRNA termination sites showing the highest level of readthrough and 144 CUT termination sites with Nrd1 (dotted) and without Nrd1 (solid). **C.** Most abundant Nrd1 motif and the most abundant Nab3 motif in the 150 nt 5' of the snoRNA and CUT terminators, respectively. MEME input parameters were any number of motifs of length 4-15 nt. Results are presented using WebLogo (Crooks et al., 2004).

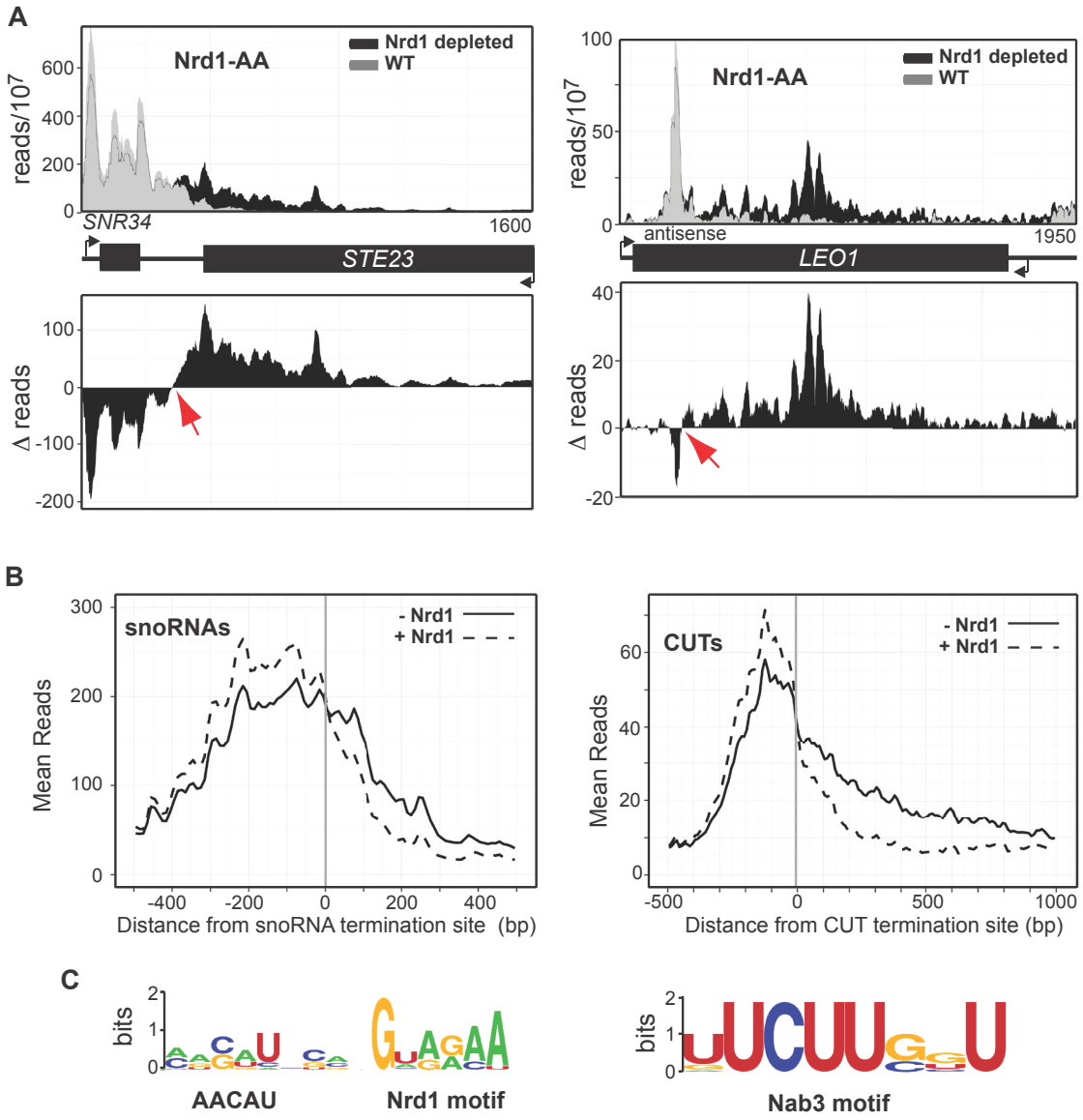

A
Nrd1-AA
Nrd1 depleted
WT
reads/10^7
600
400
200
0
SNR34
1600
STE23
100
0
-100
-200
Δ reads
Nrd1-AA
Nrd1 depleted
WT
reads/10^7
100
50
0
antisense
1950
LEO1
40
20
0
-20
Δ reads
B
snoRNAs
- Nrd1
+ Nrd1
Mean Reads
300
200
100
0
-400
-200
0
200
400
Distance from snoRNA termination site (bp)
CUTs
- Nrd1
+ Nrd1
Mean Reads
60
40
20
0
-500
0
500
1000
Distance from CUT termination site (bp)
C
bits
2
1
0
AACAU
Nrd1 motif
bits
2
1
0
Nab3 motif

Figure 8. Schulz termination vs PAR-CLIP inflection points

Relative Position on Schulz et al. termination points compared to our PAR-CLIP determined inflection points (Schulz, 2013).

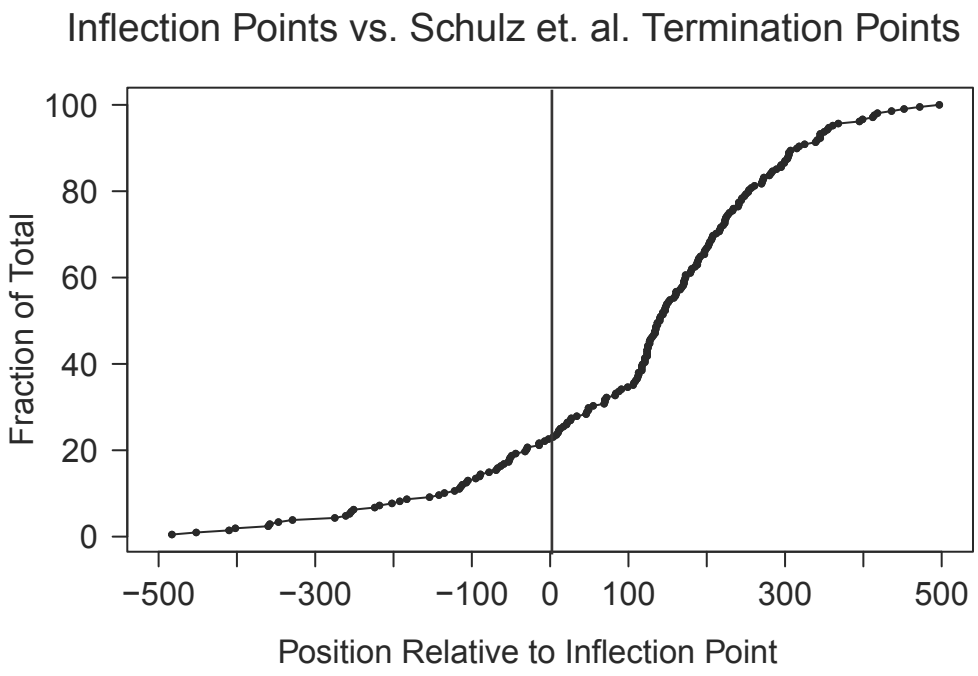
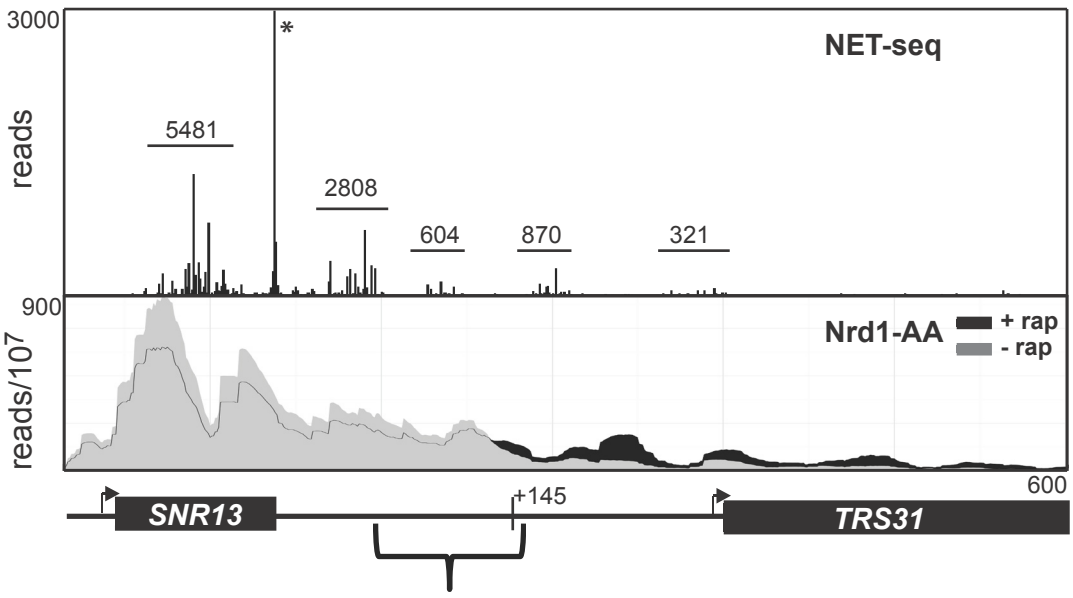

Inflection Points vs. Schulz et. al. Termination Points
Fraction of Total
100
80
60
40
20
0
−500
−300
−100
0
100
300
500
Position Relative to Inflection Point

Figure 9. SNR13 termination region

NET-Seq data from Churchman and Weissman (Churchman and Weissman, 2011) are shown above PAR-CLIP data from the *SNR13-TRS31* locus. The number of NET-seq reads from peaks of Pol II are shown above the lines. The asterisk indicates reads derived from mature snR13 RNA that contaminates the NET-seq library. The calculated *SNR13* termination region has been expanded to show U-rich sequences (red) surrounding the termination point. The sequences of several antisense CUT termination regions are shown below the SNR13 sequence.



termination region

snR13

+145

AGUACGUUGGGGUUUGGCUUGUUCAUUUUUUAUUAUCCUUUAUUUACUUGU

CUTs

+276

COG8 UGCAUUUGUAGGUAUUUUGUUUUCCCUUUUUGAAGUCCGAAUCUUUUUG

antisense

+41

LEO1 CCUCGAUGACCGCAACCCUUCUUCUUUUUUGUCUCGCUCCUAUCUUUUUC

antisense

+43

NOP15 GACUUGGCUAAUUUUUUUUAUCCUUUCUUCAUGUUUUUUGCUUCAUGUUAU

antisense

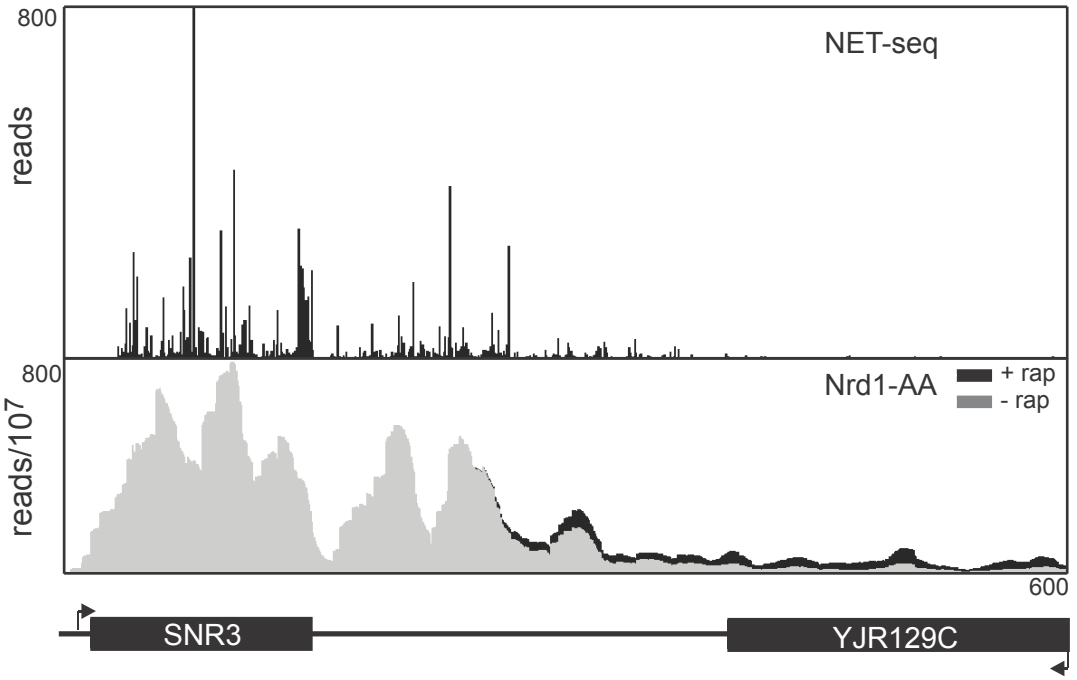
+226

BRN1 ACGUUUUUUCUUUAAUCUUCUUACGUCAACUUUUUUUCGAAACUCUUGAAUA

antisense

Figure 10. Comparison of PAR-CLIP and NET-Seq.

A comparison of NET-Seq data from Churchman and Weissman (Churchman and Weissman, 2011) to our normalized reads at the SNR3 and SNR34 locus.


800
reads
NET-seq
800
reads/10^7
Nrd1-AA
+ rap
- rap
600
SNR3
YJR129C

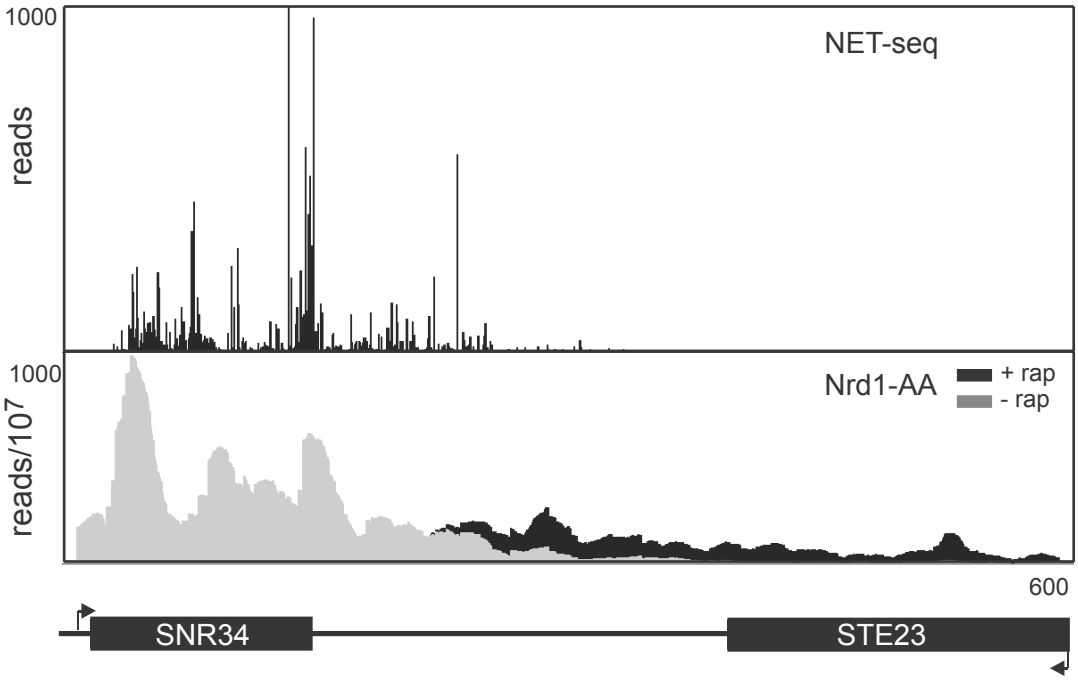

1000
reads
NET-seq
1000
reads/10^7
Nrd1-AA
+ rap
- rap
600
SNR34
STE23

Figure 11. Nrd1 increases the stability of the readthrough transcripts.

A. Northern blot and qPCR at different times after addition of rap using the amplicon highlighted in red. **B.** Histograms representing normalized reads with Nrd1 (grey) and without Nrd1 (black) at the given genomic locations. The Y-axis has been changed to emphasize the differences between treatment with (black) and without rap (grey) for the *SNR13-TRS31* locus. **C.** Similar to A but Schultz et. al. (Schulz et al., 2013) 4tU-seq data is also represented in as a histogram below each graph for the same region of the genome with Nrd1 (grey) and without Nrd1 (black).

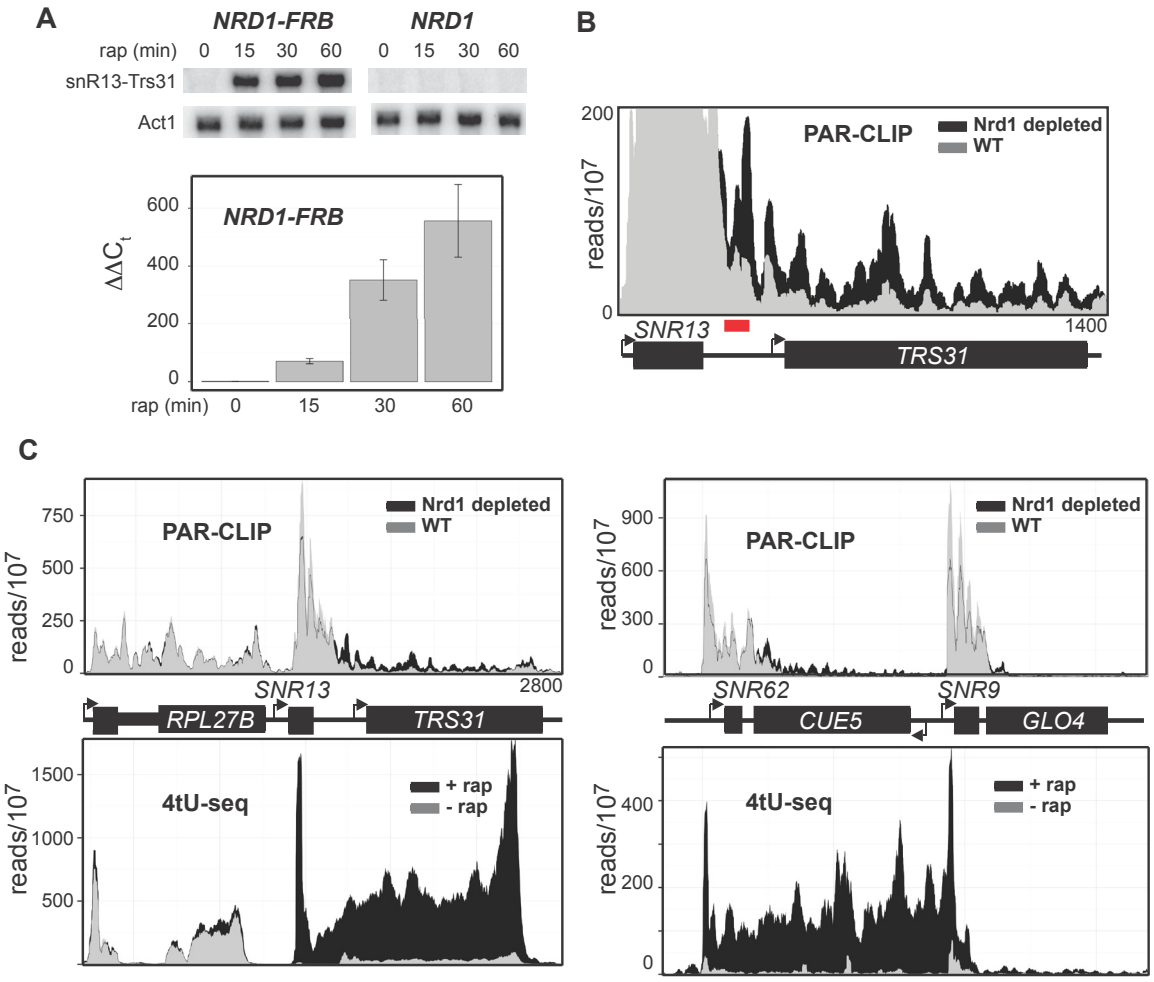

A
NRD1-FRB
NRD1
rap (min) 0 15 30 60 0 15 30 60
snR13-Trs31
Act1
NRD1-FRB
ΔΔCt
600
400
200
0
rap (min) 0 15 30 60
B
PAR-CLIP
Nrd1 depleted
WT
reads/10^7
200
0
SNR13
1400
TRS31
C
PAR-CLIP
Nrd1 depleted
WT
reads/10^7
750
500
250
0
SNR13
2800
RPL27B
TRS31
4tU-seq
+ rap
- rap
1500
1000
500
PAR-CLIP
Nrd1 depleted
WT
900
600
300
0
SNR62
SNR9
CUE5
GLO4
4tU-seq
+ rap
- rap
400
200
0

Figure 12. **Comparison of PAR-CLIP data to Schulz et. al. 4tU-seq data.**

Number of reads for each base pair at the given locus is represented by a histogram. Genes and introns are represented under each graph.

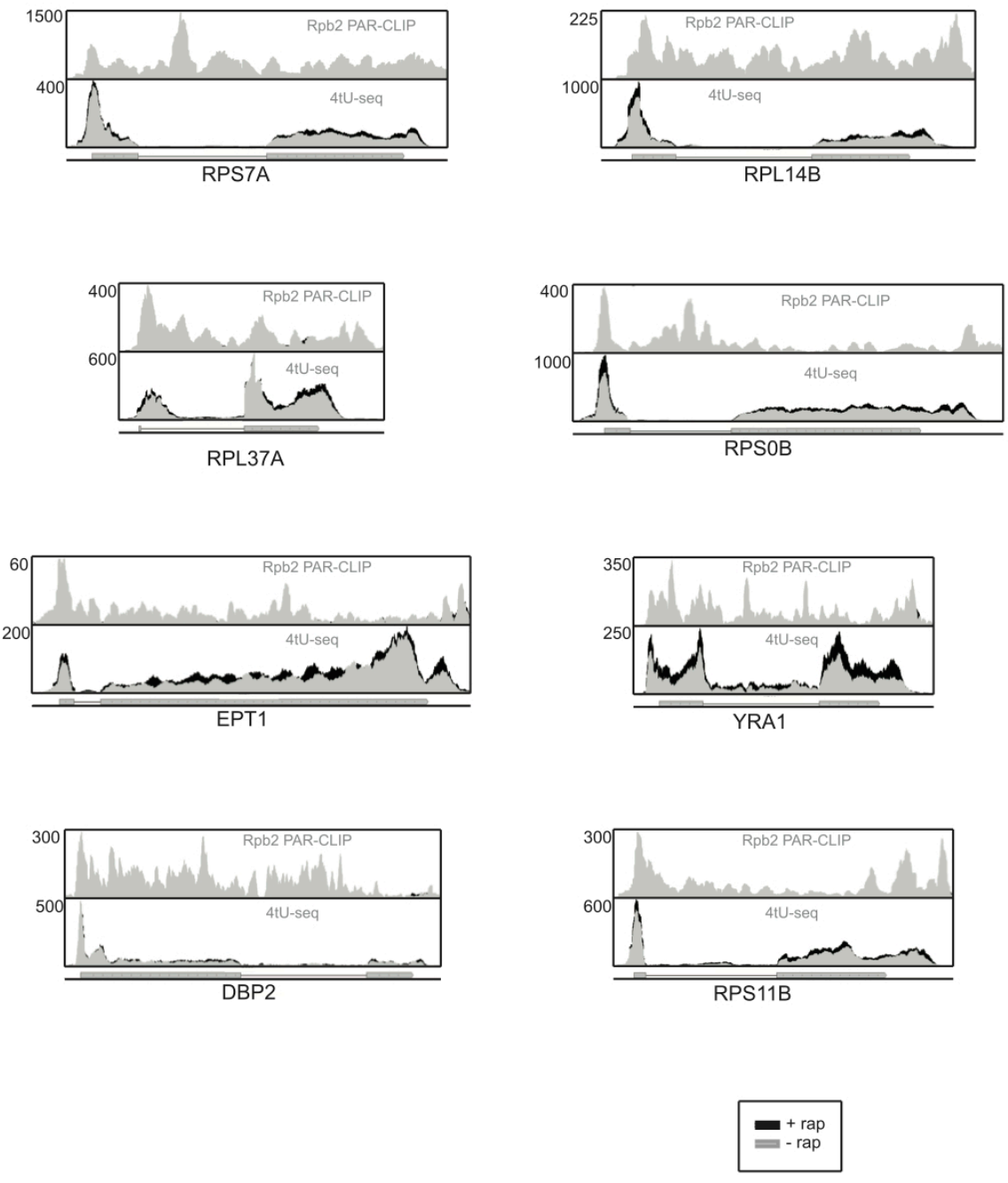

1500
Rpb2 PAR-CLIP
400
4tU-seq
RPS7A
225
Rpb2 PAR-CLIP
1000
4tU-seq
RPL14B
400
Rpb2 PAR-CLIP
600
4tU-seq
RPL37A
400
Rpb2 PAR-CLIP
1000
4tU-seq
RPS0B
60
Rpb2 PAR-CLIP
200
4tU-seq
EPT1
350
Rpb2 PAR-CLIP
250
4tU-seq
YRA1
300
Rpb2 PAR-CLIP
500
4tU-seq
DBP2
300
Rpb2 PAR-CLIP
600
4tU-seq
RPS11B
+ rap
- rap

Figure 13. Sen1 depletion causes readthrough at non-pA terminators.

A.-C. Mean reads every 10 bp for the top 500 polyA,144 calculated CUT termination sites, and 49 snoRNA termination sites, with Sen1 (dotted) and without Sen1 (solid). **D.** Histograms representing normalized reads with Sen1 (grey) and without Sen1 (black) at the given genomic locations. The same region is also represented in as a histogram below each graph with Nrd1 (grey) and without Nrd1 (black). **E.** Percent readthrough of non-pA terminators (ordered from highest to lowest) represented as a fraction of the total number of calculated non-pA terminators. *SEN1-FRB* is represented as diamonds and *NRD1-FRB* is represented as circles.

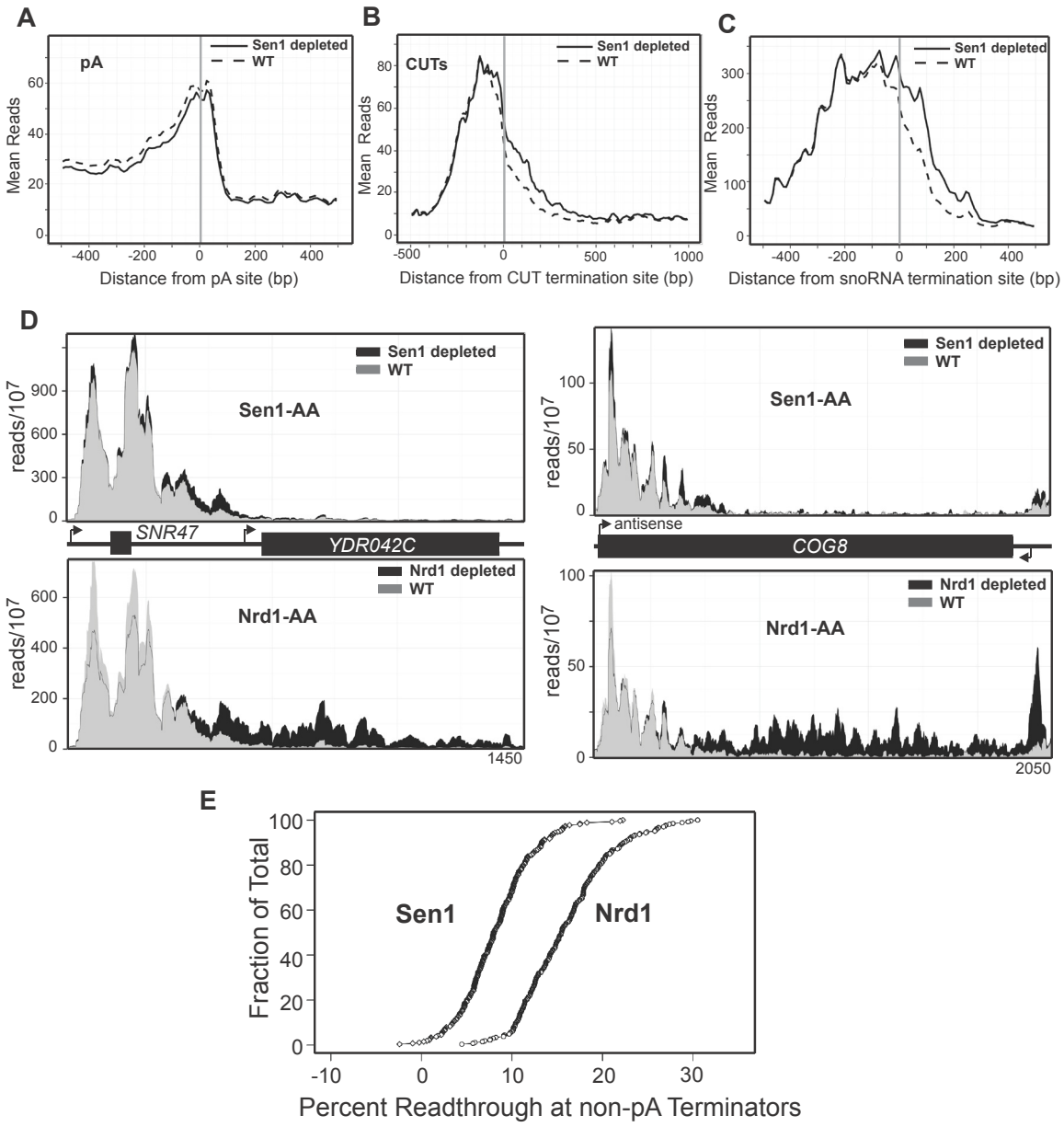

A
pA
Sen1 depleted
WT
Mean Reads
Distance from pA site (bp)
B
CUTs
Distance from CUT termination site (bp)
C
Distance from snoRNA termination site (bp)
D
Sen1-AA
reads/10^7
SNR47
YDR042C
Nrd1-AA
1450
antisense
COG8
2050
E
Fraction of Total
Sen1
Nrd1
Percent Readthrough at non-pA Terminators

Figure 14. Pol II Readthrough at Previously Described Sen1 Terminators

Histograms representing normalized reads with Sen1 (grey) and without Sen1 (black) at the given genomic locations.

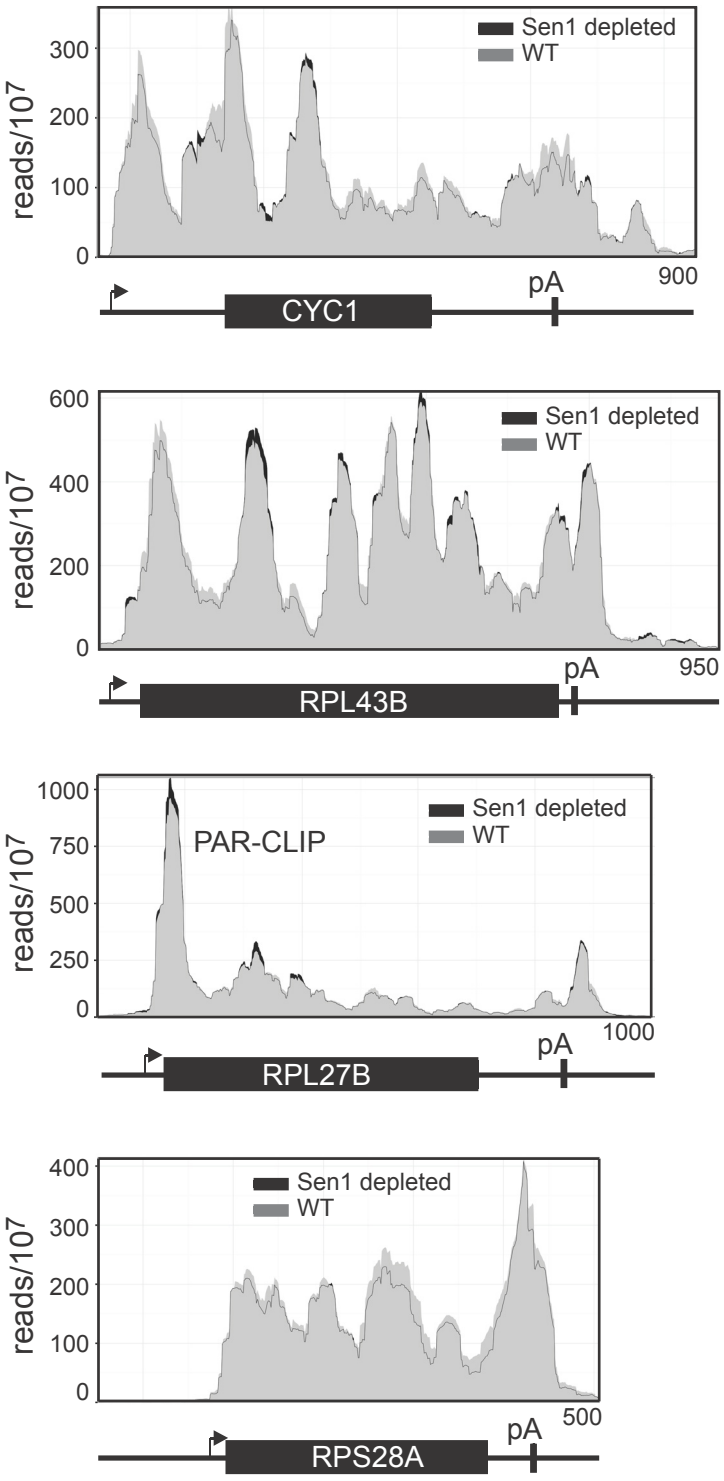

Sen1 depleted
WT
reads/10^7
300
200
100
0
900
pA
CYC1
600
400
200
0
950
pA
RPL43B
PAR-CLIP
1000
750
500
250
0
1000
pA
RPL27B
400
300
200
100
0
500
pA
RPS28A

Figure 15. NRD1 and TYE7 Pol II Readthrough.

Histograms representing normalized reads with Nrd1 (grey) and without Nrd1 (black) for NRD1 and TYE7. The same region is also represented in as a histogram below each graph with Sen1 (grey) and without Sen1 (black).

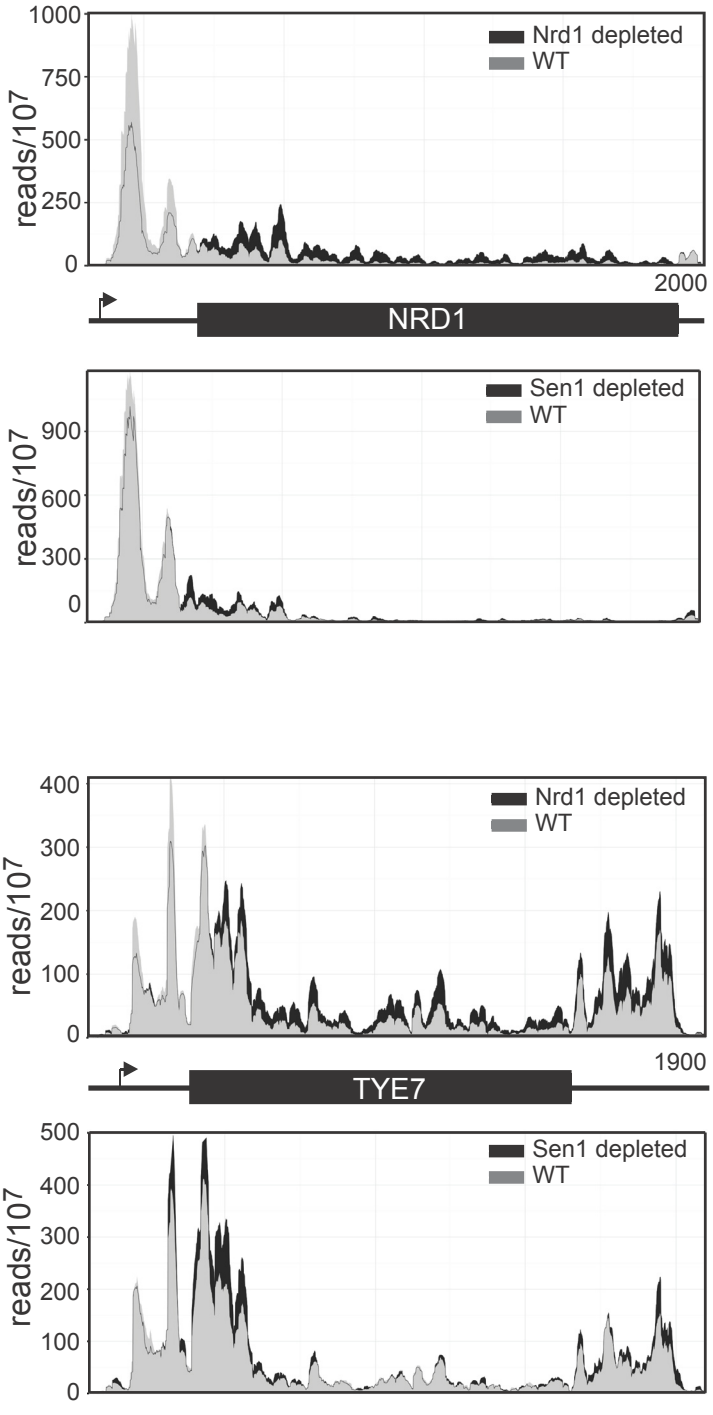

1000
750
500
250
0
reads/10^7
Nrd1 depleted
WT
2000
NRD1
900
600
300
0
Sen1 depleted
WT
400
300
200
100
0
1900
TYE7
500
400
300
200
100
0

Figure 16. Schematic representation of Pol II termination after removal of non-pA and pA termination factors.

Elongating Pol II (green) terminates pA transcripts (A) after an allosteric change (red) that reduces processivity. (B) Depletion of Ysh1 leads to minimally extended readthrough transcripts but does not block the allosteric change in Pol II. (C) Nrd1 and Nab3 binding recruit Sen1 for termination of non-pA transcripts. (D) Pol II elongation complex lacking Nrd1 does not recognize termination sequences in the nascent transcript and thus does not facilitate the allosteric transition in Pol II. This leads to processive readthrough. (E) Nrd1 and Nab3 recognize terminator sequences allowing the allosteric change in Pol II but depletion of Sen1 blocks removal of Pol II from the template.

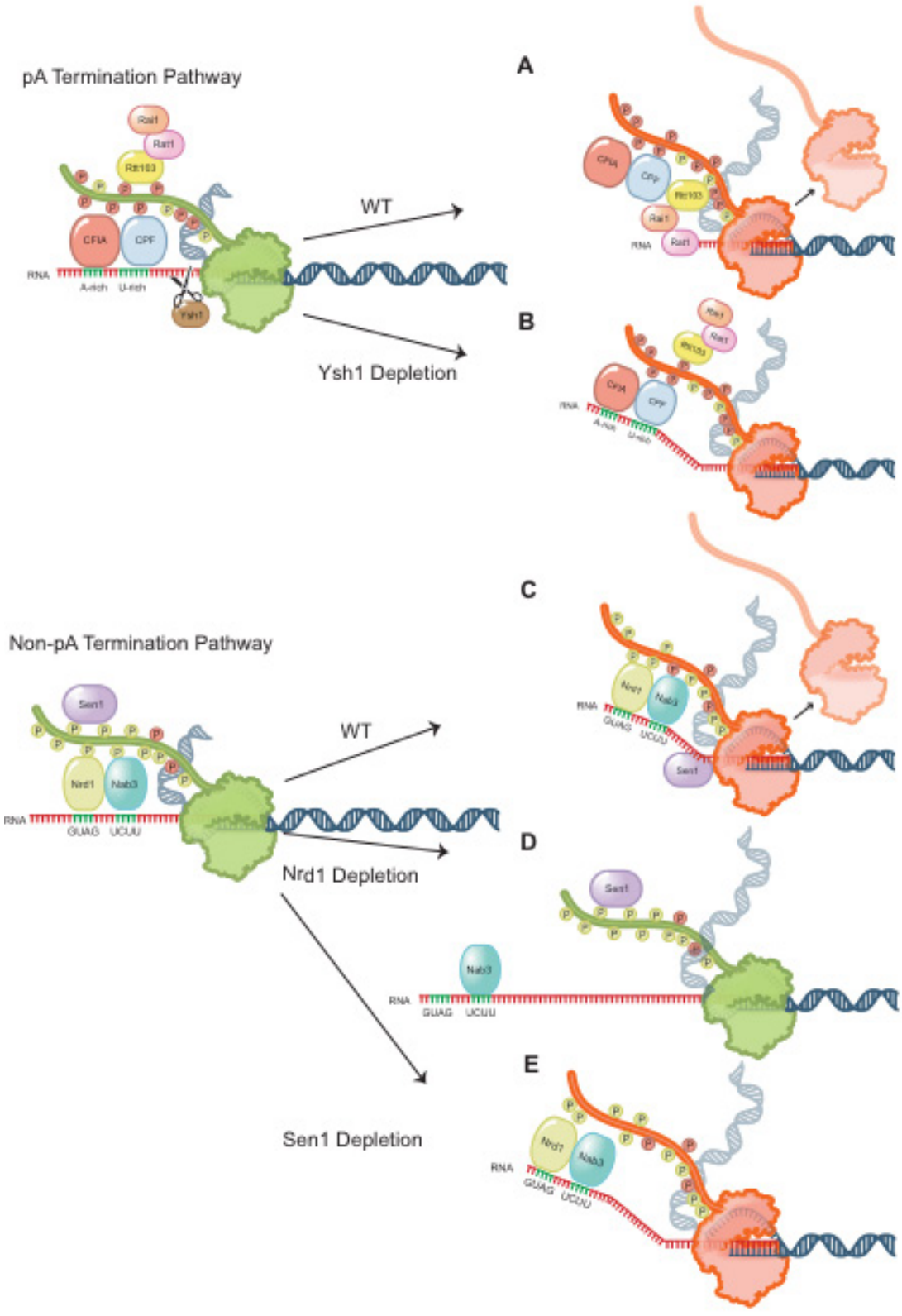

pA Termination Pathway
Rat1
Rai1
Rtt103
CFIA
CPF
RNA
A-rich
U-rich
Ysh1
WT
Ysh1 Depletion
A
B
Non-pA Termination Pathway
Sen1
Nrd1
Nab3
GUAG
UCUU
WT
Nrd1 Depletion
Sen1 Depletion
C
D
E

Table 1. Read counts at various stages of bioinformatic manipulation.

	Nrd1-FRB No Rap (Rep1)	Nrd1-FRB No Rap (Rep1)	Nrd1-FRB No Rap (Combined)
Raw Reads	78817512	89584625	171893715
Trimmed	75809544	87718484	166859712
>13 nt	55389403	66756371	126620299
Condensed	22990719	28344180	46645619
Minus tRNA	22178216	27425904	45299374
Aligned	17006647	20434186	33658009
	Nrd1-FRB + Rap (Rep1)	Nrd1-FRB + Rap (Rep1)	Nrd1-FRB + Rap (Combined)
Raw Reads	100217809	71675906	171893715
Trimmed	96033282	70826429	166859712
>13 nt	68816191	57804108	126620299
Condensed	28612700	27542687	50820528
Minus tRNA	27647347	26668871	49380255
Aligned	22001244	20797542	38392790
	Rpb2-HTB + Rap (Rep1)	Rpb2-HTB + Rap (Rep2)	Rpb2-HTB + Rap (Combined)
Raw Reads	75632597	99725061	175357658
Trimmed	69966057	95588867	165554925
>13 nt	57047117	81421645	138468763
Condensed	16021159	39400635	51683523
Minus tRNA	15240806	37967585	49877284
Aligned	10801779	25094891	33159690
	Sen1-FRB + Rap (Rep1)	Sen1-FRB + Rap (Rep2)	Sen1-FRB + Rap (Combined)
Raw Reads	48603070	132044799	180647869
Trimmed	45985758	130514127	176499885
>13 nt	37037257	103264505	140301762
Condensed	15193050	16971804	29837214
Minus tRNA	14452225	16144468	28565544
Aligned	10287141	10724525	19358671
	Ysh1-FRB + Rap		
Raw Reads	110553844		
Trimmed	105351627		
>13 nt	78948277		
Condensed	35065857		
Minus tRNA	33958529		
Aligned	23348177		

Table 2. **Top 266 points of inflection after spline function refinement.**

CUT annotations are defined in Neil et. al.(Neil, 2009). NUT annotations are defined in Schulz et. al.(Schulz, 2013). Gene and snoRNA annotations come from the Saccharomyces genome database. Top 144 CUTs highlighted in green. NA is no annotation.

Name	Chromosome	Inflection Point	Strand	Percentage Readthrough
CUT382	chrXV	553216	+	43.35
CUT721	chrXI	493135	-	40.80
CUT532	chrIV	1300554	-	39.82
NUT0165	chrIV	1365872	+	38.28
CUT285	chrXIII	130024	+	38.02
CUT088	chrIV	1240442	+	36.75
CUT301	chrXIII	535192	+	34.15
SNR65	chrIII	177261	+	34.10
CUT776	chrXIII	545526	-	33.50
CUT881	chrXVI	117824	-	32.08
CUT160	chrVII	884685	+	31.89
CUT710	chrXI	274705	-	31.84
CUT515	chrIV	892127	-	31.41
CUT110	chrV	262699	+	30.59
CUT216	chrX	582732	+	30.10
CUT075	chrIV	974762	+	29.77
NUT0377	chrX	41790	+	29.51
CUT029	chrII	576557	+	29.13
CUT144	chrVII	484779	+	28.88
CUT645	chrVIII	238387	-	28.61
CUT711	chrXI	286132	-	28.45
NUT0410	chrX	466837	+	28.18
CUT804	chrXIV	185131	-	28.16
CUT620	chrVII	762007	-	27.77
CUT300	chrXIII	501221	+	27.67
CUT114	chrV	348057	+	27.62
CUT335	chrXIV	417868	+	27.51
CUT650	chrVIII	381120	-	27.30
CUT827	chrXV	167567	-	27.00
CUT443	chrII	42768	-	26.68
CUT428	chrXVI	676079	+	26.66
CUT422	chrXVI	475789	+	26.56
CUT913	chrXVI	759202	-	26.55
CUT453	chrII	408165	-	26.53
CUT303	chrXIII	557655	+	26.48
CUT596	chrVII	337383	-	26.40
CUT695	chrX	535071	-	26.38
CUT175	chrVIII	171066	+	26.15
CUT249	chrXII	185193	+	26.08
CUT035	chrII	698509	+	25.96
CUT133	chrVII	229364	+	25.88
CUT271	chrXII	738386	+	25.82
NUT0217	chrV	563133	+	25.80
CUT720	chrXI	387150	-	25.58
NUT1393	chrXV	127850	-	25.51
CUT612	chrVII	667501	-	25.29
CUT723	chrXI	554354	-	25.26
NUT0482	chrXII	66941	+	25.13

CUT600	chrVII	412769	-	25.00
CUT507	chrIV	703891	-	24.95
CUT808	chrXIV	359108	-	24.93
NUT0447	chrXI	165231	+	24.88
CUT189	chrIX	203808	+	24.82
CUT232	chrXI	335622	+	24.81
NUT0245	chrVII	79377	+	24.57
NUT0828	chrI	224782	-	24.46
CUT825	chrXV	146911	-	24.37
NUT1010	chrVII	94513	-	24.31
CUT396	chrXV	1001332	+	24.23
NUT1273	chrXII	855366	-	24.18
SNR56	chrII	88276	+	23.99
CUT134	chrVII	269569	+	23.99
NRD1	chrXIV	174307	-	23.95
CUT817	chrXIV	529360	-	23.92
CUT781	chrXIII	646742	-	23.81
CUT109	chrV	238837	+	23.70
NUT0473	chrXI	576393	+	23.47
CUT739	chrXII	505509	-	23.24
CUT011	chrII	306412	+	23.18
CUT542	chrV	15700	-	22.95
CUT615	chrVII	702270	-	22.92
NUT1284	chrXIII	46247	-	22.80
SNR60	chrX	349046	-	22.53
CUT363	chrXV	204881	+	22.37
CUT002	chrI	138767	+	22.23
CUT919	chrXVI	874679	-	22.17
CUT742	chrXII	657775	-	22.16
SNR47	chrIV	541515	-	22.14
NUT1194	chrX	611827	-	22.07
NUT0685	chrXV	136420	+	22.07
CUT472	chrIV	99170	-	22.05
CUT241	chrXI	490117	+	21.94
CUT699	chrX	650898	-	21.83
NA	chrXI	142916	-	21.82
NUT0422	chrX	620770	+	21.78
CUT207	chrX	186970	+	21.78
NUT0344	chrVIII	294313	+	21.68
CUT655	chrVIII	450388	-	21.63
CUT095	chrIV	1359350	+	21.24
CUT415	chrXVI	352836	+	21.15
CUT880	chrXVI	108077	-	21.09
CUT019	chrII	366035	+	21.09
NUT0472	chrXI	566363	+	20.97
CUT319	chrXIV	168487	+	20.90
CUT536	chrIV	1436562	-	20.79
CUT704	chrXI	96084	-	20.71
NUT1351	chrXIV	355951	-	20.69

NUT0728	chrXV	617357	+	20.47
CUT566	chrV	416334	-	20.46
CUT082	chrIV	1149859	+	20.45
CUT332	chrXIV	349044	+	20.26
CUT112	chrV	326283	+	19.92
CUT653	chrVIII	441765	-	19.86
CUT439	chrI	35789	-	19.82
CUT724	chrXI	557207	-	19.81
CUT865	chrXV	849041	-	19.67
CUT617	chrVII	739472	-	19.66
CUT225	chrXI	31140	+	19.40
CUT813	chrXIV	459163	-	19.35
NA	chrXIII	751946	+	19.32
NUT0906	chrIV	599243	-	19.25
CUT505	chrIV	690295	-	19.09
CUT067	chrIV	588279	+	19.04
CUT669	chrIX	381386	-	19.03
CUT701	chrXI	24572	-	18.99
ARO4	chrII	717921	-	18.97
MSG5	chrXIV	530104	+	18.88
CUT081	chrIV	1106256	+	18.86
SNR45	chrXVI	822033	+	18.86
CUT575	chrVI	118993	-	18.82
NUT0645	chrXIV	307589	+	18.81
CUT049	chrIV	57222	+	18.69
CUT113	chrV	343708	+	18.63
CUT883	chrXVI	131958	-	18.62
SNR49	chrXIV	716354	+	18.60
CUT673	chrX	99146	-	18.56
CUT098	chrIV	1427703	+	18.48
CUT065	chrIV	542766	+	18.46
NUT0846	chrII	533166	-	18.40
NA	chrII	235801	-	18.37
MTD1	chrXI	590459	+	18.17
NUT0521	chrXII	500843	+	18.14
SNR34	chrXII	899484	+	18.13
NUT0097	chrIV	264737	+	18.13
NUT0352	chrVIII	461842	+	18.11
CUT092	chrIV	1319619	+	18.06
SNR4	chrV	425029	+	18.04
CUT270	chrXII	721871	+	17.92
CUT806	chrXIV	275679	-	17.92
NUT1180	chrX	380653	-	17.89
NUT0869	chrIII	310523	-	17.82
CUT278	chrXII	1042064	+	17.82
NUT0971	chrV	251351	-	17.76
NA	chrXII	827893	-	17.70
NUT0511	chrXII	388840	+	17.70
NUT1125	chrVIII	551396	-	17.68

NUT0826	chrI	184550	-	17.47
CUT267	chrXII	675014	+	17.46
CUT607	chrVII	576304	-	17.43
NA	chrVII	974893	+	17.42
CUT420	chrXVI	454235	+	17.40
SNR46	chrVII	545607	+	17.31
SNR11	chrXIII	652536	+	17.25
SNR87	chrXI	431085	-	17.24
CUT545	chrV	59103	-	17.15
CUT475	chrIV	134699	-	17.15
NUT0379	chrX	74270	+	17.12
NUT0451	chrXI	200721	+	17.04
CUT869	chrXV	953672	-	16.98
SNR62	chrXV	409626	-	16.91
NUT0356	chrVIII	554353	+	16.90
NUT0288	chrVII	767149	+	16.87
CUT587	chrVII	131130	-	16.79
NUT0107	chrIV	387686	+	16.75
CUT405	chrXVI	170206	+	16.73
SNR8	chrXV	832515	+	16.65
SNR35	chrXV	759204	-	16.58
SNR32	chrVIII	381790	+	16.57
CUT259	chrXII	404950	+	16.47
CUT705	chrXI	158166	-	16.44
CUT183	chrVIII	501773	+	16.41
CUT528	chrIV	1269536	-	16.40
CUT188	chrIX	188137	+	16.33
CUT213	chrX	413916	+	16.29
CUT550	chrV	178282	-	16.26
NUT0532	chrXII	645536	+	16.23
NUT0893	chrIV	365847	-	16.18
CUT805	chrXIV	241142	-	16.16
NUT0699	chrXV	343227	+	16.15
SNR48	chrVII	609769	+	15.96
CUT905	chrXVI	622923	-	15.91
CUT340	chrXIV	568559	+	15.88
NUT0558	chrXIII	13774	+	15.85
NA	chrVII	559366	+	15.84
CUT823	chrXV	101061	-	15.84
CUT812	chrXIV	434412	-	15.70
NUT0741	chrXV	930774	+	15.68
NUT1077	chrVII	894104	-	15.64
NUT0896	chrIV	443965	-	15.62
CUT260	chrXII	416075	+	15.55
CUT571	chrV	544874	-	15.55
NUT0160	chrIV	1297331	+	15.53
SNR10	chrVII	346294	+	15.25
CUT214	chrX	458662	+	15.21
CUT614	chrVII	671740	-	15.20

CUT399	chrXVI	41047	+	15.15
TYE7	chrXV	978041	-	15.14
NA	chrVIII	385235	+	15.09
NUT0855	chrII	786570	-	15.07
NUT0616	chrXIII	784645	+	15.02
NUT0322	chrVIII	68055	+	14.90
CUT861	chrXV	781423	-	14.89
CUT064	chrIV	502101	+	14.87
NUT1387	chrXV	79594	-	14.85
SNR71	chrVIII	411471	+	14.81
SNR83	chrXIII	626668	+	14.74
KTR2	chrXI	557877	+	14.65
NUT0256	chrVII	320259	+	14.59
GLT1	chrIV	155619	-	14.57
SNR61	chrXII	794287	-	14.54
NUT0053	chrII	615689	+	14.46
CUT374	chrXV	422133	+	14.38
CUT178	chrVIII	198778	+	14.37
NUT1523	chrXVI	882679	-	14.36
HUT1	chrXVI	87231	-	14.32
CUT051	chrIV	163549	+	14.31
NUT1384	chrXV	39983	-	14.22
CUT921	chrXVI	878303	-	14.16
NUT0959	chrV	115713	-	14.01
NUT0841	chrII	391186	-	13.97
CUT066	chrIV	563451	+	13.91
NUT1081	chrVII	958783	-	13.79
CUT744	chrXII	672513	-	13.76
NUT0932	chrIV	1183201	-	13.65
CUT015	chrII	326655	+	13.65
NUT0570	chrXIII	242291	+	13.60
CUT684	chrX	267983	-	13.60
CUT585	chrVII	55979	-	13.38
NUT0243	chrVII	64053	+	13.22
NUT0249	chrVII	173558	+	13.09
ERT1	chrII	699851	-	12.95
SNR13	chrIV	1403164	+	12.79
NUT0361	chrIX	136596	+	12.77
SNR128	chrX	139538	-	12.55
CUT368	chrXV	316180	+	12.50
CUT354	chrXV	22266	+	12.40
OLA1	chrII	291733	-	12.29
NUT0463	chrXI	446937	+	12.24
CUT733	chrXII	265028	-	11.80
NUT0117	chrIV	506605	+	11.75
SNR3	chrX	664097	+	11.73
NUT0942	chrIV	1321655	-	11.50
NUT1325	chrXIII	807785	-	11.48
NUT0363	chrIX	202534	+	11.46

CUT468	chrIV	22206	-	11.42
SNR30	chrXII	199445	+	11.36
SNR69	chrXI	365002	+	11.35
SNR31	chrXV	841952	-	11.29
NUT0218	chrVI	16922	+	11.27
CUT418	chrXVI	432564	+	11.27
CUT680	chrX	172782	-	11.04
CUT644	chrVIII	231468	-	10.98
KAP95	chrXII	823355	-	10.87
SNR36	chrXV	680685	-	10.85
NUT0861	chrIII	124701	-	10.70
SNR42	chrXI	558935	-	10.54
NUT1114	chrVIII	455034	-	10.48
NA	chrXI	317694	+	10.25
NUT0763	chrXVI	95857	+	10.24
NUT0554	chrXII	1002486	+	10.11
SNR82	chrVII	317156	+	9.75
NUT0310	chrVII	1060500	+	9.63
SNR63	chrIV	323229	-	9.01
CUT775	chrXIII	513189	-	7.96
SNR5	chrXV	842699	+	7.81
CUT901	chrXVI	481060	-	4.09

Table 3. **Top 49 points of inflection 3' of snoRNAs after spline function refinement.**

snoRNA annotations come from the saccharomyces genome database.

Name	Chromosome	Inflection Point	Strand	Percentage Readthrough
SNR34	chrXII	899483	+	23.42
SNR47	chrIV	541516	-	22.94
SNR65	chrIII	177256	+	20.01
SNR45	chrXVI	822025	+	18.55
SNR4	chrV	425034	+	18.12
SNR56	chrII	88276	+	18.01
SNR49	chrXIV	716317	+	17.44
SNR13	chrIV	1403160	+	17.26
SNR60	chrX	349128	-	17.11
SNR32	chrVIII	381790	+	17.04
SNR35	chrXV	759197	-	16.58
SNR62	chrXV	409624	-	15.68
SNR10	chrVII	346326	+	15.59
SNR46	chrVII	545603	+	15.42
SNR71	chrVIII	411468	+	15.19
SNR48	chrVII	609780	+	13.17
SNR82	chrVII	317151	+	12.87
SNR64	chrXI	39050	+	12.83
SNR61	chrXII	794281	-	12.82
SNR8	chrXV	832499	+	12.71
SNR83	chrXIII	626531	+	12.65
SNR31	chrXV	841956	-	12.38
SNR128	chrX	139535	-	12.18
SNR79	chrXII	348332	-	12.03
SNR3	chrX	664089	+	11.94
SNR30	chrXII	199426	+	11.84
SNR36	chrXV	680692	-	11.58
SNR66	chrXIV	586314	+	11.34
SNR69	chrXI	364951	+	11.10
SNR42	chrXI	558934	-	10.65
SNR63	chrIV	323224	-	10.36
SNR58	chrXV	136035	-	9.97
SNR5	chrXV	842694	+	9.74
SNR85	chrXIII	67783	-	9.71
SNR39B	chrVII	366381	-	9.13
SNR43	chrIII	107553	-	8.94
SNR37	chrX	228003	-	8.48
SNR40	chrXIV	89478	+	8.48
SNR87	chrXI	431096	-	8.17
SNR59	chrXVI	173793	+	7.97
SNR67	chrV	61594	+	7.70
SNR161	chrII	307132	-	7.59
SNR53	chrV	61707	+	7.44
SNR50	chrXV	259670	+	7.12
SNR189	chrIII	178587	-	6.87
SNR33	chrIII	142316	-	6.64
SNR81	chrXV	234528	+	5.86
SNR17A	chrXV	780654	+	5.44
SNR18	chrI	143034	+	2.60

Table 4. Yeast strains used in this study.

Yeast Strain	Genotype
HHY168	MATalpha tor1-1 fpr1::NAT RPL13A-2×FKBP12::TRP1
JC2749	MATalpha tor1-1 fpr1::NAT RPL13A-2×FKBP12::TRP1 RPB2-HTB::KanMX
JC2771	MATalpha tor1-1 fpr1::NAT RPL13A-2×FKBP12::TRP1 RPB2-HTB::KanMX NRD1-FRB::HIS3
JC2807	MATalpha tor1-1 fpr1::NAT RPL13A-2×FKBP12::TRP1 RPB2-HTB::KanMX NRD1-FRB::HIS3 /pRS406-NRD1
JC2794	MATalpha tor1-1 fpr1::NAT RPL13A-2×FKBP12::TRP1 RPB2-HTB::KanMX NRD1-FRBGFP::HIS3 RPB3-TAGRFP::hph
JC2775	MATalpha tor1-1 fpr1::NAT RPL13A-2×FKBP12::TRP1 RPB2-HTB::KanMX YSH1-FRB::HIS3
JC2770	MATalpha tor1-1 fpr1::NAT RPL13A-2×FKBP12::TRP1 RPB2-HTB::KanMX SEN1-FRB::HIS3
JC2828	MATalpha tor1-1 fpr1::NAT RPL13A-2×FKBP12::TRP1 YSH1-FRBGFP::HIS3 RPB3-TAGRFP::hph
JC2834	MATalpha tor1-1 fpr1::NAT RPL13A-2×FKBP12::TRP1 <KanMX>PgalSEN1-FRBGFP::HIS3 RPB3-TAGRFP::hph

Table 5. **Primers used in this study.**

Primer Name	Sequence
3' Adaptor	AppAGATCGGAAGAGCACACGTCTddC
5' Adaptor	GUUCAGAGUUCUACAGUCCGACGAUC
RT Primer	AGACGTGTGCTCTTCCGATCT
PCR Amp Fwd	AATGATACGGCGACCACCGAGATCTACACGTTGAGAGTTCT ACAGTCCG*A
PCR Amp Rev 1	CAAGCAGAAGACGGCATACGAGATATTGGCGTGACTGGAG TTCAGACGTGTGCTCTTCGGATC*T
PCR Amp Rev 2	CAAGCAGAAGACGGCATACGAGATTACAAGGTGACTGGAG TTCAGACGTGTGCTCTTCCGATC*T
snr13 qPCR/Northern Fwd	CCTGTCTTTCTGTCGCTTCC
snr13 qPCR Rev	AGTACGTTGGGTTTGGCTTG
ACT1 qPCR/Northern Fwd	GCCTTGGACTTCGAACAAGA
ACT1 qPCR Rev	CAAAACCCAAAACAGAAGGA
RPL42A qPCR Fwd	GGGTGGTGAAAAGAAGCAAA
RPL42A qPCR Rev	TGCAACTGCTTGTCTGGAAA
RPL34A qPCR Fwd	GTCAAGGAACAAACCGAAGC
RPL34A qPCR Rev	ATTCCCATGGCTATGGATGA
RPL21B qPCR Fwd	CTCCAGTTCCATACGAAACCTT
RPL21B qPCR Rev	TGAATACCTGCTGAGAAGATGG
RPL24B qPCR Fwd	CGCCACTTCCCGTTAAGATA
RPL24B qPCR Rev	CGGGTAACGATCATGAACAA

(where a * denotes and phosphorothioate bond)

REFERENCES

Al Husini, N., Kudla, P., and Ansari, A. (2013). A role for CF1A 3' end processing complex in promoter-associated transcription. PLoS Genet *9*, e1003722.

Arigo, J.T., Carroll, K.L., Ames, J.M., and Corden, J.L. (2006a). Regulation of yeast NRD1 expression by premature transcription termination. Mol Cell *21*, 641-651.

Arigo, J.T., Eyler, D.E., Carroll, K.L., and Corden, J.L. (2006b). Termination of cryptic unstable transcripts is directed by yeast RNA-binding proteins Nrd1 and Nab3. Mol Cell *23*, 841-851.

Bacikova, V., Pasulka, J., Kubicek, K., and Stefl, R. (2014). Structure and semi-sequence-specific RNA binding of Nrd1. Nucleic Acids Res.

Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W., and Noble, W.S. (2009). MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res *37*, W202-208.

Barta, I., and Iggo, R. (1995). Autoregulation of expression of the yeast Dbp2p 'DEAD-box' protein is mediated by sequences in the conserved DBP2 intron. EMBO J *14*, 3800-3808.

Berretta, J., and Morillon, A. (2009). Pervasive transcription constitutes a new level of eukaryotic genome regulation. EMBO Rep *10*, 973-982.

Birse, C.E., Minvielle-Sebastia, L., Lee, B.A., Keller, W., and Proudfoot, N.J. (1998). Coupling termination of transcription to messenger RNA maturation in yeast. Science *280*, 298-301.

Brannan, K., and Bentley, D.L. (2012). Control of Transcriptional Elongation by RNA Polymerase II: A Retrospective. Genet Res Int *2012*, 170173.

Brendolise, C., Rouillard, J.M., Dufour, M.E., and Lacroute, F. (2002). Expression analysis of RNA14, a gene involved in mRNA 3' end maturation in yeast: characterization of the rna14-5 mutant strain. Mol Genet Genomics *267*, 515-525.

Brow, D.A. (2011). Sen-sing RNA Terminators. Mol Cell *42*, 717-718.

Buratowski, S. (2005). Connections between mRNA 3' end processing and transcription termination. Curr Opin Cell Biol *17*, 257-261.

Burger, K., Muhl, B., Kellner, M., Rohrmoser, M., Gruber-Eber, A., Windhager, L., Friedel, C.C., Dolken, L., and Eick, D. (2013). 4-thiouridine inhibits rRNA synthesis and causes a nucleolar stress response. RNA Biol *10*.

Calvo, O., and Manley, J.L. (2001). Evolutionarily conserved interaction between CstF-64 and PC4 links transcription, polyadenylation, and termination. Mol Cell *7*, 1013-1023.

Carroll, K.L., Ghirlando, R., Ames, J.M., and Corden, J.L. (2007). Interaction of yeast RNA-binding proteins Nrd1 and Nab3 with RNA polymerase II terminator elements. RNA *13*, 361-373.

Carroll, K.L., Pradhan, D.A., Granek, J.A., Clarke, N.D., and Corden, J.L. (2004). Identification of cis elements directing termination of yeast nonpolyadenylated snoRNA transcripts. Mol Cell Biol *24*, 6241-6252.

Castelnuovo, M., Rahman, S., Guffanti, E., Infantino, V., Stutz, F., and Zenklusen, D. (2013). Bimodal expression of PHO84 is modulated by early termination of antisense transcription. Nat Struct Mol Biol *20*, 851-858.

Chinchilla, K., Rodriguez-Molina, J.B., Ursic, D., Finkel, J.S., Ansari, A.Z., and Culbertson, M.R. (2012). Interactions of Sen1, Nrd1, and Nab3 with multiple phosphorylated forms of the Rpb1 C-terminal domain in Saccharomyces cerevisiae. Eukaryot Cell *11*, 417-429.

Churchman, L.S., and Weissman, J.S. (2011). Nascent transcript sequencing visualizes transcription at nucleotide resolution. Nature *469*, 368-373.

Cloutier, S.C., Ma, W.K., Nguyen, L.T., and Tran, E.J. (2012). The DEAD-box RNA helicase Dbp2 connects RNA quality control with repression of aberrant transcription. J Biol Chem *287*, 26155-26166.

Cloutier, S.C., Wang, S., Ma, W.K., Petell, C.J., and Tran, E.J. (2013). Long noncoding RNAs promote transcriptional poising of inducible genes. PLoS Biol *11*, e1001715.

Connelly, S., and Manley, J.L. (1988). A functional mRNA polyadenylation signal is required for transcription termination by RNA polymerase II. Genes Dev *2*, 440-452.

Conrad, N.K., Wilson, S.M., Steinmetz, E.J., Patturajan, M., Brow, D.A., Swanson, M.S., and Corden, J.L. (2000). A yeast heterogeneous nuclear ribonucleoprotein complex associated with RNA polymerase II. Genetics *154*, 557-571.

Creamer, T.J., Darby, M.M., Jamonnak, N., Schaughency, P., Hao, H., Wheelan, S.J., and Corden, J.L. (2011). Transcriptome-wide binding sites for components of the Saccharomyces cerevisiae non-poly(A) termination pathway: Nrd1, Nab3, and Sen1. PLoS Genet *7*, e1002329.

Crooks, G.E., Hon, G., Chandonia, J.M., and Brenner, S.E. (2004). WebLogo: a sequence logo generator. Genome Res *14*, 1188-1190.

Davis, C.A., and Ares, M., Jr. (2006). Accumulation of unstable promoter-associated transcripts upon loss of the nuclear exosome subunit Rrp6p in Saccharomyces cerevisiae. Proc Natl Acad Sci U S A *103*, 3262-3267.

Fan, X., Moqtaderi, Z., Jin, Y., Zhang, Y., Liu, X.S., and Struhl, K. (2010). Nucleosome depletion at yeast terminators is not intrinsic and can occur by a transcriptional mechanism linked to 3'-end formation. Proc Natl Acad Sci U S A *107*, 17945-17950.

Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J.*, et al.* (2004). Bioconductor: open software development for computational biology and bioinformatics. Genome Biol *5*, R80.

Gromak, N., West, S., and Proudfoot, N.J. (2006). Pause sites promote transcriptional termination of mammalian RNA polymerase II. Mol Cell Biol *26*, 3986-3996.

Grzechnik, P., Tan-Wong, S.M., and Proudfoot, N.J. (2014). Terminate and make a loop: regulation of transcriptional directionality. Trends Biochem Sci.

Gudipati, R.K., Xu, Z., Lebreton, A., Seraphin, B., Steinmetz, L.M., Jacquier, A., and Libri, D. (2012). Extensive degradation of RNA precursors by the exosome in wild-type cells. Mol Cell *48*, 409-421.

Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano, M., Jr., Jungkamp, A.C., Munschauer, M.*, et al.* (2010a). Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. Cell *141*, 129-141.

Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano, M., Jungkamp, A.C., Munschauer, M.*, et al.* (2010b). PAR-CliP--a method to identify transcriptome-wide the binding sites of RNA binding proteins. J Vis Exp.

Haruki, H., Nishikawa, J., and Laemmli, U.K. (2008). The anchor-away technique: rapid, conditional establishment of yeast mutant phenotypes. Mol Cell *31*, 925-932.

Hsin, J.P., and Manley, J.L. (2012). The RNA polymerase II CTD coordinates transcription and RNA processing. Genes Dev *26*, 2119-2137.

Huang, Y., Weng, X., and Russu, I.M. (2010). Structural energetics of the adenine tract from an intrinsic transcription terminator. J Mol Biol *397*, 677-688.

Hyman, L.E., and Moore, C.L. (1993). Termination and pausing of RNA polymerase II downstream of yeast polyadenylation sites. Mol Cell Biol *13*, 5159-5167.

Jacquier, A. (2009). The complex eukaryotic transcriptome: unexpected pervasive transcription and novel small RNAs. Nat Rev Genet *10*, 833-844.

Jamonnak, N., Creamer, T.J., Darby, M.M., Schaughency, P., Wheelan, S.J., and Corden, J.L. (2011). Yeast Nrd1, Nab3, and Sen1 transcriptome-wide binding maps suggest multiple roles in post-transcriptional RNA processing. RNA *17*, 2011-2025.

Jensen, T.H., Jacquier, A., and Libri, D. (2013). Dealing with pervasive transcription. Mol Cell *52*, 473-484.

Kazerouninia, A., Ngo, B., and Martinson, H.G. (2010). Poly(A) signal-dependent degradation of unprocessed nascent transcripts accompanies poly(A) signal-dependent transcriptional pausing in vitro. RNA *16*, 197-210.

Kim, H.D., Choe, J., and Seo, Y.S. (1999). The sen1(+) gene of Schizosaccharomyces pombe, a homologue of budding yeast SEN1, encodes an RNA and DNA helicase. Biochemistry *38*, 14697-14710.

Kim, M., Ahn, S.H., Krogan, N.J., Greenblatt, J.F., and Buratowski, S. (2004a). Transitions in RNA polymerase II elongation complexes at the 3' ends of genes. EMBO J *23*, 354-364.

Kim, M., Krogan, N.J., Vasiljeva, L., Rando, O.J., Nedea, E., Greenblatt, J.F., and Buratowski, S. (2004b). The yeast Rat1 exonuclease promotes transcription termination by RNA polymerase II. Nature *432*, 517-522.

Kim, M., Vasiljeva, L., Rando, O.J., Zhelkovsky, A., Moore, C., and Buratowski, S. (2006). Distinct pathways for snoRNA and mRNA termination. Mol Cell *24*, 723-734.
Kireeva, M.L., Komissarova, N., Waugh, D.S., and Kashlev, M. (2000). The 8-nucleotide-long RNA:DNA hybrid is a primary stability determinant of the RNA polymerase II elongation complex. J Biol Chem *275*, 6530-6536.

Kuehner, J.N., and Brow, D.A. (2008). Regulation of a eukaryotic gene by GTP-dependent start site selection and transcription attenuation. Mol Cell *31*, 201-211.

Kuehner, J.N., Pearson, E.L., and Moore, C. (2011). Unravelling the means to an end: RNA polymerase II transcription termination. Nat Rev Mol Cell Biol *12*, 283-294.

Langmead, B. (2010). Aligning short sequencing reads with Bowtie. Curr Protoc Bioinformatics *Chapter 11*, Unit 11 17.

Larson, D.R., Zenklusen, D., Wu, B., Chao, J.A., and Singer, R.H. (2011). Real-time observation of transcription initiation and elongation on an endogenous yeast gene. Science *332*, 475-478.

Logan, J., Falck-Pedersen, E., Darnell, J.E., Jr., and Shenk, T. (1987). A poly(A) addition site and a downstream termination region are required for efficient

cessation of transcription by RNA polymerase II in the mouse beta maj-globin gene. Proc Natl Acad Sci U S A *84*, 8306-8310.

Luo, W., and Bentley, D. (2004). A ribonucleolytic rat torpedoes RNA polymerase II. Cell *119*, 911-914.

Luo, W., Johnson, A.W., and Bentley, D.L. (2006). The role of Rat1 in coupling mRNA 3'-end processing to transcription termination: implications for a unified allosteric-torpedo model. Genes Dev *20*, 954-965.

Ma, W.K., Cloutier, S.C., and Tran, E.J. (2013). The DEAD-box protein Dbp2 functions with the RNA-binding protein Yra1 to promote mRNP assembly. J Mol Biol *425*, 3824-3838.

Mandart, E. (1998). Effects of mutations in the Saccharomyces cerevisiae RNA14 gene on the abundance and polyadenylation of its transcripts. Mol Gen Genet *258*, 16-25.

Marquardt, S., Hazelbaker, D.Z., and Buratowski, S. (2011). Distinct RNA degradation pathways and 3' extensions of yeast non-coding RNA species. Transcription *2*, 145-154.

Martin, F.H., and Tinoco, I., Jr. (1980). DNA-RNA hybrid duplexes containing oligo(dA:rU) sequences are exceptionally unstable and may facilitate termination of transcription. Nucleic Acids Res *8*, 2295-2299.

Mischo, H.E., and Proudfoot, N.J. (2013). Disengaging polymerase: terminating RNA polymerase II transcription in budding yeast. Biochim Biophys Acta *1829*, 174-185.

Moqtaderi, Z., Geisberg, J.V., Jin, Y., Fan, X., and Struhl, K. (2013). Species-specific factors mediate extensive heterogeneity of mRNA 3' ends in yeasts. Proc Natl Acad Sci U S A *110*, 11073-11078.

Nag, A., Narsinh, K., Kazerouninia, A., and Martinson, H.G. (2006). The conserved AAUAAA hexamer of the poly(A) signal can act alone to trigger a stable decrease in RNA polymerase II transcription velocity. RNA *12*, 1534-1544.

Nag, A., Narsinh, K., and Martinson, H.G. (2007). The poly(A)-dependent transcriptional pause is mediated by CPSF acting on the body of the polymerase. Nat Struct Mol Biol *14*, 662-669.

Neil, H., Malabat, C., d'Aubenton-Carafa, Y., Xu, Z., Steinmetz, L.M., and Jacquier, A. (2009). Widespread bidirectional promoters are the major source of cryptic transcripts in yeast. Nature *457*, 1038-1042.

Ozsolak, F., Kapranov, P., Foissac, S., Kim, S.W., Fishilevich, E., Monaghan, A.P., John, B., and Milos, P.M. (2010). Comprehensive polyadenylation site maps in yeast and human reveal pervasive alternative polyadenylation. Cell *143*, 1018-1029.

Pearson, E.L., and Moore, C.L. (2013). Dismantling promoter-driven RNA polymerase II transcription complexes in vitro by the termination factor Rat1. J Biol Chem *288*, 19750-19759.

Pelechano, V., Chavez, S., and Perez-Ortin, J.E. (2010). A complete set of nascent transcription rates for yeast genes. PLoS One *5*, e15442.

Porrua, O., Hobor, F., Boulay, J., Kubicek, K., D'Aubenton-Carafa, Y., Gudipati, R.K., Stefl, R., and Libri, D. (2012). In vivo SELEX reveals novel sequence and structural determinants of Nrd1-Nab3-Sen1-dependent transcription termination. EMBO J *31*, 3935-3948.

Porrua, O., and Libri, D. (2013). A bacterial-like mechanism for transcription termination by the Sen1p helicase in budding yeast. Nat Struct Mol Biol *20*, 884-891.

Rahl, P.B., Lin, C.Y., Seila, A.C., Flynn, R.A., McCuine, S., Burge, C.B., Sharp, P.A., and Young, R.A. (2010). c-Myc regulates transcriptional pause release. Cell *141*, 432-445.

Rines, D.R., Thomann, D., Dorn, J.F., Goodwin, P., and Sorger, P.K. (2011). Live cell imaging of yeast. Cold Spring Harb Protoc *2011*.

Schreieck, A., Easter, A.D., Etzold, S., Wiederhold, K., Lidschreiber, M., Cramer, P., and Passmore, L.A. (2014). RNA polymerase II termination involves C-terminal-domain tyrosine dephosphorylation by CPF subunit Glc7. Nat Struct Mol Biol *21*, 175-179.

Schulz, D., Schwalb, B., Kiesel, A., Baejen, C., Torkler, P., Gagneur, J., Soeding, J., and Cramer, P. (2013). Transcriptome Surveillance by Selective Termination of Noncoding RNA Synthesis. Cell.

Shearwin, K.E., Callen, B.P., and Egan, J.B. (2005). Transcriptional interference--a crash course. Trends Genet *21*, 339-345.

Steinmetz, E.J., and Brow, D.A. (1996). Repression of gene expression by an exogenous sequence element acting in concert with a heterogeneous nuclear ribonucleoprotein-like protein, Nrd1, and the putative helicase Sen1. Mol Cell Biol *16*, 6993-7003.

Steinmetz, E.J., and Brow, D.A. (1998). Control of pre-mRNA accumulation by the essential yeast protein Nrd1 requires high-affinity transcript binding and a domain implicated in RNA polymerase II association. Proc Natl Acad Sci U S A *95*, 6699-6704.

Steinmetz, E.J., Conrad, N.K., Brow, D.A., and Corden, J.L. (2001). RNA-binding protein Nrd1 directs poly(A)-independent 3'-end formation of RNA polymerase II transcripts. Nature *413*, 327-331.

Steinmetz, E.J., Ng, S.B., Cloute, J.P., and Brow, D.A. (2006a). cis- and trans-Acting determinants of transcription termination by yeast RNA polymerase II. Mol Cell Biol *26*, 2688-2696.

Steinmetz, E.J., Warren, C.L., Kuehner, J.N., Panbehi, B., Ansari, A.Z., and Brow, D.A. (2006b). Genome-wide distribution of yeast RNA polymerase II and its control by Sen1 helicase. Mol Cell *24*, 735-746.

Tagwerker, C., Zhang, H., Wang, X., Larsen, L.S., Lathrop, R.H., Hatfield, G.W., Auer, B., Huang, L., and Kaiser, P. (2006). HB tag modules for PCR-based gene tagging and tandem affinity purification in Saccharomyces cerevisiae. Yeast *23*, 623-632.

Tan-Wong, S.M., Zaugg, J.B., Camblong, J., Xu, Z., Zhang, D.W., Mischo, H.E., Ansari, A.Z., Luscombe, N.M., Steinmetz, L.M., and Proudfoot, N.J. (2012). Gene loops enhance transcriptional directionality. Science *338*, 671-675.

Thiebaut, M., Colin, J., Neil, H., Jacquier, A., Seraphin, B., Lacroute, F., and Libri, D. (2008). Futile cycle of transcription initiation and termination modulates the response to nucleotide shortage in S. cerevisiae. Mol Cell *31*, 671-682.

Thiebaut, M., Kisseleva-Romanova, E., Rougemaille, M., Boulay, J., and Libri, D. (2006). Transcription termination and nuclear degradation of cryptic unstable transcripts: a role for the nrd1-nab3 pathway in genome surveillance. Mol Cell *23*, 853-864.

Ursic, D., Chinchilla, K., Finkel, J.S., and Culbertson, M.R. (2004). Multiple protein/protein and protein/RNA interactions suggest roles for yeast DNA/RNA helicase Sen1p in transcription, transcription-coupled DNA repair and RNA processing. Nucleic Acids Res *32*, 2441-2452.

Vasiljeva, L., and Buratowski, S. (2006). Nrd1 interacts with the nuclear exosome for 3' processing of RNA polymerase II transcripts. Mol Cell *21*, 239-248.

Vasiljeva, L., Kim, M., Mutschler, H., Buratowski, S., and Meinhart, A. (2008). The Nrd1-Nab3-Sen1 termination complex interacts with the Ser5-phosphorylated RNA polymerase II C-terminal domain. Nat Struct Mol Biol *15*, 795-804.

West, S., Gromak, N., and Proudfoot, N.J. (2004). Human 5' --> 3' exonuclease Xrn2 promotes transcription termination at co-transcriptional cleavage sites. Nature *432*, 522-525.

Wilkening, S., Pelechano, V., Jarvelin, A.I., Tekkedil, M.M., Anders, S., Benes, V., and Steinmetz, L.M. (2013). An efficient method for genome-wide polyadenylation site mapping and RNA quantification. Nucleic Acids Res *41*, e65.

Wlotzka, W., Kudla, G., Granneman, S., and Tollervey, D. (2011). The nuclear RNA polymerase II surveillance system targets polymerase III transcripts. EMBO J *30*, 1790-1803.

Wyers, F., Rougemaille, M., Badis, G., Rousselle, J.C., Dufour, M.E., Boulay, J., Regnault, B., Devaux, F., Namane, A., Seraphin, B., *et al.* (2005). Cryptic pol II transcripts are degraded by a nuclear quality control pathway involving a new poly(A) polymerase. Cell *121*, 725-737.

Xu, Z., Wei, W., Gagneur, J., Perocchi, F., Clauder-Munster, S., Camblong, J., Guffanti, E., Stutz, F., Huber, W., and Steinmetz, L.M. (2009). Bidirectional promoters generate pervasive transcription in yeast. Nature *457*, 1033-1037.

Zhang, Z., and Gilmour, D.S. (2006). Pcf11 is a termination factor in Drosophila that dismantles the elongation complex by bridging the CTD of RNA polymerase II to the nascent transcript. Mol Cell *21*, 65-74.

Zhang, Z., Klatt, A., Henderson, A.J., and Gilmour, D.S. (2007). Transcription termination factor Pcf11 limits the processivity of Pol II on an HIV provirus to repress gene expression. Genes Dev *21*, 1609-1614.

Zhelkovsky, A., Tacahashi, Y., Nasser, T., He, X., Sterzer, U., Jensen, T.H., Domdey, H., and Moore, C. (2006). The role of the Brr5/Ysh1 C-terminal domain and its homolog Syc1 in mRNA 3'-end processing in Saccharomyces cerevisiae. RNA *12*, 435-445.

Part 2: Mapping RBM16 to the Mammalian Transcriptome

Paul Schaughency and Jeffry L. Corden

ABSTRACT

The C-terminal Domain (CTD) of RNA Polymerase II (Pol II) is a major part of a regulatory network that links transcription to RNA processing. In *Saccharomyces cerevisiae,* Nrd1 is a transcriptional factor that binds to the CTD and nascent RNA and terminates transcription on a variety of specific RNAs. Nrd1 shares homology to the proteins RBM16 and SFRS15 in mammalian cells. Photoactivatable-Ribonucleoside-Enhanced Crosslinking and Immunopreciptiation (PAR-CLIP) has been used to define the RNAs that are bound to human RBM16. Of the PAR-CLIP products that were purified, non-coding RNAs (ncRNAs), mRNAs, and snoRNAs were the highly represented. This observation the possibility of a shared function between RBM16 and Nrd1.

INTRODUCTION

RNA Polymerase II (Pol II) is responsible for the transcription of most protein coding genes and some ncRNAs in eukaryotes. The C-terminal domain (CTD) of Pol II is a scaffold to which many proteins bind during each of a transcription cycle (Buratowski, 2009). The CTD itself is modified to reflect the state of Pol II transcription (Chapman et al., 2007; Zhang and Corden, 1991). It is through the "CTD code" of phosphorylation states and other post-translational modifications that different factors and complexes can bind to influence transcription (Carroll et al., 2007; Nedea et al., 2003).

In yeast, the Nrd1-Nab3-Sen1 (NNS) pathway is a complex of proteins involved in the attenuation of certain Pol II transcripts (Steinmetz et al., 2001). Nrd1 is a protein that interacts with CTD through its CTD interacting domain (CID) and the nascent RNA through its RNA recognition motif (RRM). These two domains define the function of this protein. Using homology and looking for a similar protein architecture two other related proteins were discovered in higher eukaryotes. These proteins were first referred to as SCAF4 and SCAF8, but now they are referred to as RBM16 and SFRS15 respectfully (Yuryev et al., 1996). Crystallization of the CID from these proteins seems to indicate that they interact with the CTD in a very similar manner (Lunde et al., 2010).

The NNS complex, in yeast, terminates the transcripts of snoRNAs, some mRNAS, and cryptic unstable transcripts (CUTs) (Arigo et al., 2006a; Arigo et al., 2006b; Steinmetz et al., 2001). Similar to yeast, snoRNA maturation requires the nuclear exosome (Berndt et al., 2012) Although an NNS-like pathway has never been found in higher eukaryotes, this similarity might suggest a mechanistic conservation. Unlike yeast,

C/D box snoRNAs in mammals are mostly found in the introns of genes (Hirose et al., 2003). This functional difference means that although there may be a similarity between Nrd1 and RBM16 on the protein level the mechanism of action would have to be different if RBM16 is involved in snoRNA maturation.

RBM16's function within the nucleus of mammalian cells is poorly understood. Using PAR-CLIP, we have pulled down and purified RNAs that are bound to RBM16. Definition of the RNAs purified, might give insight into the possible functions of this protein and how it compares to its yeast counterpart Nrd1. Further evidence of similarity between types of RNAs bound to both of these factors might translate to a similar mechanism of transcriptional regulation.

RESULTS

Expression of 6xHis-TEV-Biotin tagged version of RBM16 was done using the Invitrogen flp-in Trex system. Correct expression of the tagged RBM16 was verified by western blot after treatment with doxycycline for 6, 12, 24, and 48 hours (Figure 17A). After 12 hours of doxycycline treatment, expression of RBM16-HTB was approximately equal to that of endogenous RBM16 levels (Figure 17A). For cross-linking exeriments, 4-thiouridine and doxycycline were both added to the cell culture media and allowed to incubate with the cells for 12 hours.

A single cDNA library preparation was made from the RNA crosslinked to purified of RBM16-HTB after treatment. A 6% polyacrylamide nucleic acid gel showed a plethora of PCR amplified cDNA from the purification (Figure 17B). Bioanalyzer results indicated that the final cDNA libarary was between 100-150 bps long (Figure 17C). Notice, that there was minor contamination by primer dimers even after the gel extraction.

Of the 32,923,767 sequences derived from sequencing the library, only 25,165,211 sequences had an identifiable 3' adaptor. After condensing the reads to remove duplicates 5,402,948 reads were left, of which 367,700 reads aligned to the hg19 genome only twice. This is only 6.8 % of the condensed reads and 1.1 % of the total reads before trimming.

Within those reads that did align, 273 genes were enriched within the dataset (Table 6). Of note are ncRNAs: Xist, NCRNA00164, and MALAT1 and several protein coding genes: HFM1, PDE3A, SLIT2, AUTS2, and SRRM2. Of those genes that had aligned reads, most had more reads at the 3' end of the gene than that 5' end of the gene

(Figure 18). 59 snoRNA loci were also enriched in this dataset (Table 2). Most of these snoRNAs were C/D box snoRNAs.

DISCUSSION

Expression of a tagged version of RBM16 was the first step to purification and elucidation of target RNAs. The spliced gene for RBM16 is approximately 4kb in length. Using Invitrogen's Flp-in Trex expression system, we were able to incorporate a tetracycline/doxycyline inducible HTB tagged version of RBM16 into HEK-293 cells. The system uses Flp recombinase, Zeocin susceptibility, and hygromycin resistance to guarantee the protein is integrated into the genome. Further testing by western blot showed that, indeed, we could get reliable expression of our tagged RBM16 within 12 hours of treatment with doxycycline (Figure 17A).

Using a combination of PAR-CLIP and highly denaturing tandem affinity purification, we were able to purify RNAs that were specifically bound to RBM16. Even though the resulting library was sparse in terms of reads that aligned to the genome, we were able to identify some RNAs that RBM16 binds.

The largest amounts of reads aligned were of RBM16 binding to Xist (Table 6). Xist is a non-coding RNA involved in X chromosome inactivation (Maclary et al., 2013). Some other ncRNAs were found bound to RBM16 as well. Although the number of reads aligned to these ncRNAs were not as significant as Xist, this may point to a role for RBM16 in transcriptional control of ncRNAs. Although, Xist and MALAT1 seem to be pulled down in most human cell culture PAR-CLIP experiments (Friedersdorf and Keene, 2014). Other genes with high number of reads aligned were those of protein coding genes. Interestingly, of all genes that had as least 10 reads, more reads were found in the 3' end of the genes than the 5' end of the gene (Figure 18).

Of the 59 snoRNA genes that RBM16 binds, 51 are C/D box and 8 are H/ACA (Table 7). Most of these C/D box snoRNAs are contained within the introns of protein coding genes while H/ACA are processed in a different manner (Berndt et al., 2012; Hirose et al., 2003). Perhaps, RBM16 retains some function to regulate the maturation of snoRNAs. As the exosome is still conserved in the maturation of snoRNAs, RBM16 may act to bridge these two processes. This mechanism, however, would not be through transcriptional termination, but an alternative unknown mechanism most likely involving the splicing machinery.

This dataset hints at the function of RBM16 within the mammalian nucleus. In yeast, Nrd1 binds and recruits the nuclear exosome and is required for proper snoRNA maturation (Grzechnik and Kufel, 2008; Kim et al., 2006; Vasiljeva and Buratowski, 2006). It does seem through the pulldown of C/D box snoRNAs RBM16 retains some function in the maturation of snoRNA from its yeast homolog Nrd1. Nrd1 is also important in controlling pervasive transcription of non-coding RNAs and some mRNAs (Arigo et al., 2006a; Arigo et al., 2006b; Creamer, 2011; Thiebaut et al., 2006). It is possible that RBM16 is also part of some complex regulating ncRNAs protein coding genes, again, through linkage to the nuclear exosome. Although RBM16's mechanism of action could be quite different from that of Nrd1, their final outcome for transcriptional control of some RNAs might be similar. This foundation may lead to further evidence of these assumptions.

MATERIALS AND METHODS

Expression of RBM16-HTB in HEK-293 cells

Knocking in of the HTB tagged version of RBM16 was done using the Invitrogen Flp-In TRex system per the manufacturers instructions. Cells that were hygromycin resistant were tested by treating with 1ug/ml doxycycline for up to 48 hours and measuring protein levels using a western blot as previously described (Figure 17A) (Creamer, 2011).

Purification of RBM16-HTB

HEK-293 cells that expressed RBM16-HTB after doxycycline treatment were grown to approximately 90% confluency in 10 100mm cell culture dishes. Cells were treated with 1ug/ml doxycycline and 4-thiouridine for 12 hours. After treatment, the cells were irradiated using a mercury arc-lamp at 365nm for 10s. Further steps taken during purification were identical to what has been described previously (Creamer, 2011).

Library Preparation

Libraries were made as previously described (Creamer, 2011). Resulting sequences were also analyzed as previously described (Schaughency, 2014), but with the following difference: the sequences were aligned using the hg19 genome.

Figure 17. Expression and purification of RBM16-HTB

A. Western blot of whole cell extracts made from cells treated with doxycycline for the indicated times. Top western was probed with an anti-RBM16 antibody and the bottom western was probed with an anti-RGS6XHIS antibody. B. 6% polyacrylamide gel of the cDNA library after PCR amplification. C. Visual representation of a Bioanalyzer lane containing the gel purified fragment of the cDNA library.

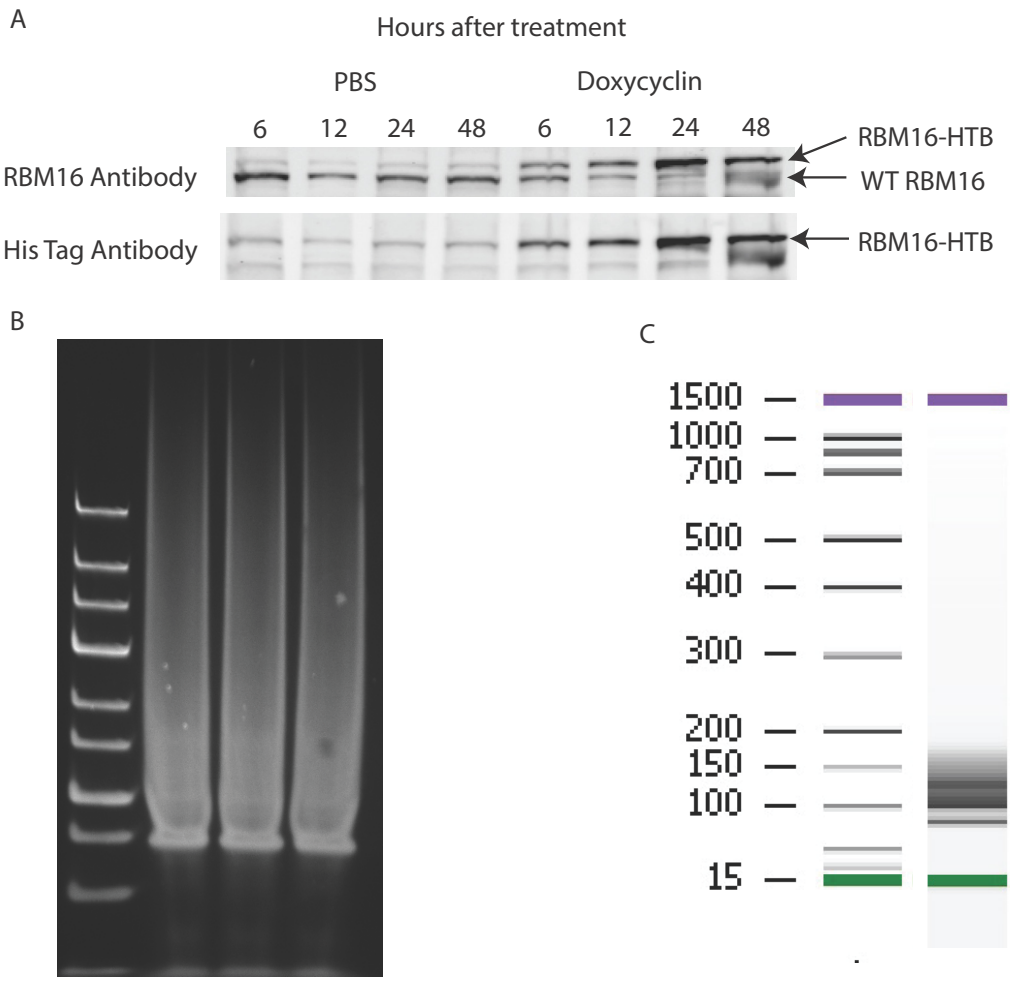

A
Hours after treatment
PBS
Doxycyclin
6
12
24
48
6
12
24
48
RBM16 Antibody
His Tag Antibody
RBM16-HTB
WT RBM16
RBM16-HTB
B
C
1500
1000
700
500
400
300
200
150
100
15

Figure 18. Reads that map to the 5' or 3' end of each gene

Pie-chart of the number of reads that were at the 3' end of a given gene (black) vs the 5' end (grey).

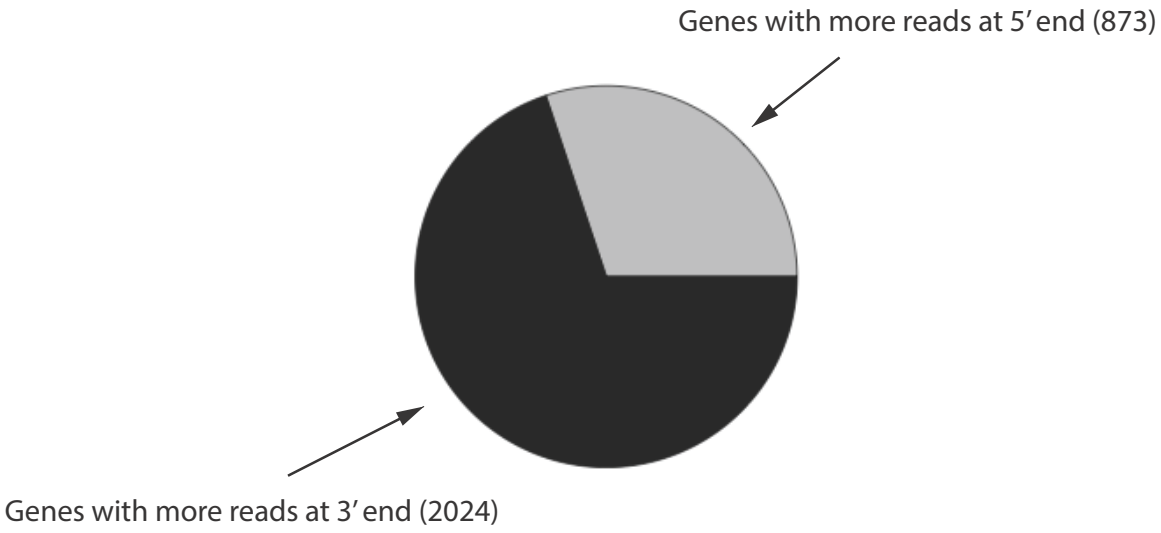

Genes with more reads at 5′ end (873)
Genes with more reads at 3′ end (2024)

Table 6. Genes with RBM16 reads aligned

List of all genes containing more than 100 reads aligned in order from highest to lowest.

Chromosome	Start	End	Gene	Reads	Strand
chrX	73040495	73072588	XIST	3051	-
chr1	91726323	91870426	HFM1	991	-
chr2	132905164	133015542	NCRNA00164	724	-
chr12	20522197	20833881	PDE3A	504	+
chr4	20255235	20620788	SLIT2	435	+
chr7	69063905	70257885	AUTS2	384	+
chr16	2802330	2821413	SRRM2	378	+
chr11	65265233	65273940	MALAT1	373	+
chr5	106712590	107006596	EFNA5	359	-
chrX	53559063	53713673	HUWE1	320	-
chr21	34915350	34949812	SON	315	+
chr16	78133551	79246564	WWOX	312	+
chr3	59735036	61237133	FHIT	296	-
chr3	61547243	62280573	PTPRG	294	+
chr10	52750945	54055274	PRKG1	292	+
chr21	38739859	38887679	DYRK1A	282	+
chr1	21132972	21503340	EIF4G3	278	-
chr8	70584575	70747208	SLCO5A1	276	-
chrX	24712056	25015103	POLA1	272	+
chr5	14143829	14509458	TRIO	267	+
chr8	48685669	48872743	PRKDC	263	-
chr6	56322785	56819413	DST	263	-
chr10	34400097	35103923	PARD3	260	-
chr1	237205702	237997288	RYR2	258	+
chr1	205681947	205719361	NUCKS1	252	-
chr6	155054512	155155194	RBM16	252	+
chr2	212240442	213403352	ERBB4	245	-
chr8	118811602	119124058	EXT1	240	-
chr4	30722037	31148421	PCDH7	235	+
chr1	245013602	245027827	HNRNPU	233	-
chr2	140988996	142889270	LRP1B	233	-
chr1	61542946	61928460	NFIA	230	+
chr7	127292202	127732659	SND1	221	+

chr12	1100404	1602697	ERC1	219	+
chr13	34392206	34540695	RFC3	216	+
chr21	32490736	32931290	TIAM1	212	-
chr21	27252861	27543446	APP	211	-
chr6	73331835	73905580	KCNQ5	209	+
chrX	53401070	53449618	SMC1A	208	-
chr19	3976054	3985461	EEF2	206	-
chr11	115044345	115375241	CADM1	204	-
chr10	129894923	129924655	MKI67	201	-
chr9	134305477	134375578	BAT2L1	200	+
chr21	47744036	47865682	PCNT	199	+
chr16	89334035	89556969	ANKRD11	197	-
chr1	171454666	171562650	BAT2L2	196	+
chr22	38879443	38902345	DDX17	195	-
chr5	179125930	179158642	CANX	194	+
chr6	16299343	16761721	ATXN1	194	-
chr22	28374002	29075853	TTC28	190	-
chr1	243651535	244006553	AKT3	190	-
chr1	214776532	214837914	CENPF	189	+
chr7	132937823	133750514	EXOC4	188	+
chr8	10753654	11058875	XKR6	184	-
chr1	35649201	35658743	SFPQ	183	-
chr3	78646390	79639061	ROBO1	182	-
chr5	89854617	90460033	GPR98	182	+
chr7	28338940	28865511	CREB5	181	+
chr6	157099086	157530402	ARID1B	180	+
chr1	19401000	19536746	UBR4	179	-
chr7	111366164	111846462	DOCK4	178	-
chr5	138609792	138666130	MATR3	178	+
chr16	58553855	58663750	CNOT1	177	-
chr4	73940502	74124502	ANKRD17	173	-
chr15	99192761	99507759	IGF1R	173	+
chr20	34291531	34330193	RBM39	172	-
chr7	3341080	4308632	SDK1	172	+

chr7	26229556	26240413	HNRNPA2B1	171	-
chr14	68286509	69062614	RAD51L1	171	+
chr1	39547118	39952789	MACF1	171	+
chr6	100956608	101329224	ASCC3	169	-
chr8	128806779	129113499	PVT1	169	+
chr5	179041179	179050722	HNRNPH1	167	-
chrX	12156585	12740529	FRMPD4	167	+
chr10	70715892	70744279	DDX21	166	+
chr10	7204249	7453450	SFMBT2	163	-
chr1	8412464	8877699	RERE	162	-
chr6	90353231	90529442	MDN1	162	-
chr19	10764937	10803095	ILF3	162	+
chr12	862225	1020618	WNK1	162	+
chr12	116396381	116714991	MED13L	161	-
chr1	233119882	233431459	PCNXL2	160	-
chr1	225117356	225586996	DNAH14	160	+
chr1	31404353	31538564	PUM1	160	-
chr7	110303110	111202347	IMMP2L	159	-
chr1	115259537	115300671	CSDE1	159	-
chr5	58264865	59189621	PDE4D	157	-
chr6	20534688	21231764	CDKAL1	157	+
chr16	70557701	70606201	SF3B3	156	+
chrX	137713734	138287185	FGF13	155	-
chr1	50906935	51425936	FAF1	155	-
chr7	98476113	98610864	TRRAP	153	+
chr6	168227671	168372703	MLLT4	152	+
chr6	124125069	125146786	NKAIN2	151	+
chr3	184032356	184053146	EIF4G1	151	+
chr15	60780483	61521502	RORA	150	-
chr21	33683330	33765312	URB1	150	-
chr17	29421995	29704695	NF1	150	+
chr21	38123189	38362503	HLCS	150	-
chrX	131760038	132095423	HS6ST2	150	-
chr17	56078280	56084707	SFRS1	150	-

chr14	102547075	102606086	HSP90AA1	149	-
chr14	102430865	102517135	DYNC1H1	149	+
chr10	5726801	5806943	C10orf18	149	+
chr17	44107282	44249594	KIAA1267	148	-
chr12	10998448	11324213	PRR4	146	-
chr6	17615269	17706818	NUP153	146	-
chr17	1553923	1588176	PRPF8	146	-
chr2	232319459	232329205	NCL	145	-
chr20	47662838	47713486	CSE1L	144	+
chr10	64926985	65225722	JMJD1C	144	-
chr2	61705069	61765418	XPO1	144	-
chr11	68228199	68382802	SAPS3	143	+
chr1	16174359	16266950	SPEN	143	+
chr12	110719032	110788898	ATP2A2	143	+
chr6	1624041	2245846	GMDS	142	-
chr6	20402137	20493945	E2F3	141	+
chr6	37787307	38122400	ZFAND3	141	+
chr5	36876861	37065921	NIPBL	141	+
chr10	114206756	114578503	VTI1A	140	+
chr13	93879078	95060274	GPC6	139	+
chr12	11033560	11324210	PRH1	139	-
chr17	27582854	27621166	NUFIP2	139	-
chr1	6845384	7829764	CAMTA1	138	+
chr7	101459292	101927250	CUX1	138	+
chrX	84498997	84528368	ZNF711	138	+
chr11	10818593	10830582	EIF4G2	137	-
chr1	91380855	91487671	ZNF644	137	-
chr5	154092462	154197165	LARP1	137	+
chr17	78518625	78940173	RPTOR	137	+
chr9	20344968	20622514	MLLT3	136	-
chr10	126676418	126849624	CTBP2	136	-
chr10	28821427	28909925	WAC	136	+
chr14	21677295	21737638	HNRNPC	135	-
chr6	128289924	128841870	PTPRK	135	-

chr6	79644136	79788011	PHIP	135	-
chr6	35800811	35888956	SRPK1	134	-
chr17	58755172	59470199	BCAS3	134	+
chr16	3775055	3930121	CREBBP	134	-
chr8	26149034	26228646	PPP2R2A	133	+
chr12	54624730	54673915	CBX5	133	-
chr19	36705504	36729676	ZNF146	133	+
chr1	197053257	197115824	ASPM	132	-
chr7	132469623	132766833	CHCHD3	132	-
chr17	64298926	64806862	PRKCA	132	+
chr4	151185594	151936879	LRBA	131	-
chr2	236402736	237034127	AGAP1	131	+
chr18	32073254	32471808	DTNA	131	+
chr16	30710462	30751450	SRCAP	129	+
chr2	32582096	32843966	BIRC6	129	+
chr11	65192517	65207155	NEAT1	129	+
chr3	173116244	174001116	NLGN1	129	+
chr17	35441927	35766902	ACACA	128	-
chr9	94972625	95056038	IARS	128	-
chr17	45727275	45761004	KPNB1	128	+
chr1	2160134	2241652	SKI	128	+
chr7	139044634	139108200	LUC7L2	128	+
chrX	70752933	70795740	OGT	128	+
chr21	35445823	35515334	MRPS6	127	+
chrX	73164159	73290219	NCRNA00183	126	+
chr20	36661948	36720766	RPRD1B	126	+
chr13	39917029	40177356	LHFP	125	-
chrX	40944888	41095832	USP9X	125	+
chr13	60239717	60738119	DIAPH3	125	-
chr21	46188955	46221738	UBE2G2	124	-
chr3	71004736	71633140	FOXP1	124	-
chr20	40031170	40247133	CHD6	124	-
chr7	151832010	152133090	MLL3	124	-
chr22	47158549	47569726	TBC1D22A	124	+

chr7	156931655	157062066	UBE3C	124	+
chr12	14518611	14651697	ATF7IP	124	+
chr5	53180614	53606403	ARL15	124	-
chr11	102980160	103350591	DYNC2H1	124	+
chr17	38544795	38574169	TOP2A	123	-
chr5	10353828	10435491	MARCH6	123	+
chr17	55333931	55757299	MSI2	123	+
chr3	9439403	9519838	SETD5	123	+
chr1	227177566	227505826	CDC42BPA	122	-
chr14	66974125	67648525	GPHN	122	+
chr5	65892176	66463092	MAST4	122	+
chr19	12810008	12834810	TNPO2	121	-
chr10	11047259	11378674	CELF2	120	+
chr21	40557404	40685556	BRWD1	120	-
chr5	145826873	145891069	TCERG1	120	+
chrX	123094475	123236506	STAG2	119	+
chr2	71558889	71662189	ZNF638	119	+
chr6	133562495	133853258	EYA4	119	+
chr5	65440085	65476716	SFRS12	118	+
chr20	61509090	61569274	DIDO1	118	-
chr6	32936437	32949282	BRD2	118	+
chr3	54156693	55108584	CACNA2D3	118	+
chr22	36134783	36424585	RBM9	118	-
chr21	30671220	30734215	BACH1	117	+
chr10	75910965	76469060	ADK	117	+
chr5	60628100	60841999	ZSWIM6	117	+
chr13	96743093	97491812	HS6ST3	116	+
chr13	30083551	30169825	SLC7A1	116	-
chr21	44263204	44299678	WDR4	116	-
chr7	155437203	155574179	RBM33	116	+
chr1	182808451	182856884	DHX9	116	+
chr5	140894588	140998622	DIAPH1	116	-
chr15	93443551	93571237	CHD2	116	+
chr1	236958581	237067281	MTR	115	+

chr4	24529087	24586184	DHX15	115	-
chr3	171757418	172118493	FNDC3B	115	+
chr8	18384813	18871196	PSD3	115	-
chr15	85923871	86292586	AKAP13	115	+
chrX	76760356	77041719	ATRX	115	-
chr3	99536681	99897447	C3orf26	114	+
chr22	33669062	34316416	LARGE	114	-
chr6	111982485	112194627	FYN	114	-
chr10	120794541	120840334	EIF3A	114	-
chr1	225674534	225840845	ENAH	114	-
chr1	236713971	236767814	HEATR1	113	-
chr2	216974020	217071016	XRCC5	113	+
chr21	34106210	34144169	C21orf66	113	-
chr7	104756823	105029341	SRPK2	111	-
chr6	18224400	18264799	DEK	111	-
chr10	114710009	114926073	TCF7L2	111	+
chr10	103544200	103578222	MGEA5	111	-
chr5	31400601	31532282	RNASEN	111	-
chr6	5261584	5771816	FARS2	111	+
chr1	116915836	116947396	ATP1A1	111	+
chr20	8113296	8865547	PLCB1	110	+
chr16	24741049	24837548	TNRC6A	110	+
chr17	8377530	8534036	MYH10	110	-
chr5	32354456	32444844	ZFR	110	-
chr16	53737875	54148381	FTO	110	+
chr3	160117430	160152741	SMC4	110	+
chr5	176560080	176727214	NSD1	110	+
chr6	97590037	97731052	C6orf167	109	-
chr2	29117533	29171080	WDR43	109	+
chr17	60019966	60142643	MED13	109	-
chr21	45432206	45526433	TRAPPC10	108	+
chr5	71403118	71505397	MAP1B	108	+
chr20	13976146	16033842	MACROD2	108	+
chr21	33043313	33104431	SFRS15	108	-

chr22	18270415	18507325	MICAL3	107	-
chr3	50712672	51421629	DOCK3	107	+
chr7	130794855	131181398	MKLN1	107	+
chr3	151985829	152183569	MBNL1	107	+
chr8	9413445	9639856	TNKS	107	+
chr10	98277867	98346809	TM9SF3	107	-
chr6	44796469	45345670	SUPT3H	106	-
chr8	17780366	17887457	PCM1	106	+
chr18	72342919	72777628	ZNF407	106	+
chr12	133200348	133263945	POLE	106	-
chr19	41768391	41813811	HNRNPUL1	106	+
chr11	118307205	118395936	MLL	105	+
chr1	89149922	89301938	PKN2	105	+
chr3	23244784	23632296	UBE2E2	105	+
chr2	54683454	54898583	SPTBN1	105	+
chr3	65339906	66024509	MAGI1	105	-
chr5	72112418	72210215	TNPO1	105	+
chr20	21283942	21370463	XRN2	105	+
chr1	78030190	78148343	ZZZ3	104	-
chr19	10244022	10305755	DNMT1	104	-
chr2	205410516	206480537	PARD3B	104	+
chr10	320130	735608	DIP2C	104	-
chr3	168867270	169381556	MECOM	104	-
chr16	18816175	18937726	SMG1	104	-
chr6	57182422	57513376	PRIM2	104	+
chr17	62494374	62502484	DDX5	103	-
chr20	49506883	49547527	ADNP	103	-
chr9	126141933	126692417	DENND1A	103	-
chr10	855484	931702	LARP4B	103	-
chr17	59759985	59940755	BRIP1	102	-
chr21	45079432	45115960	RRP1B	102	+
chr19	34663352	34720420	LSM14A	101	+
chr6	89319989	89673348	RNGTT	101	-
chr8	131064353	131414217	ASAP1	101	-

chr1	97187175	97280605	PTBP2	101	+
chr4	1873123	1983934	WHSC1	101	+

Table 7. snoRNAS with RBM16 reads aligned

List of all snoRNAs containing more that five reads aligned in order from highest to lowest.

Chromosome	Start	End	Gene	Reads	Strand
chr1	45241537	45241610	SNORD55	61	+
chr19	12817263	12817332	SNORD41	53	-
chr17	74554874	74554951	SNORD1C	34	+
chr19	49994164	49994229	SNORD34	32	+
chr19	49993874	49993956	SNORD33	32	+
chr20	2634858	2634932	SNORD110	22	+
chr19	3982505	3982570	SNORD37	20	-
chr11	122929970	122930180	SNORD14C	18	-
chr11	62622764	62622838	SNORD26	16	-
chr22	39715057	39715118	SNORD43	16	-
chr11	62622093	62622167	SNORD28	14	-
chr8	56986398	56986460	SNORD54	14	-
chr16	89627838	89627909	SNORD68	14	+
chr1	173835106	173835166	SNORD44	13	-
chr17	62223438	62223517	SNORD104	13	+
chr20	2637585	2637656	SNORD57	12	+
chr19	49994432	49994517	SNORD35A	12	+
chr1	173836812	173836883	SNORD74	12	-
chr1	28833877	28834083	SNORA73A	11	+
chr2	232325079	232325153	SNORD82	11	-
chrX	135961358	135961430	SNORD61	11	-
chr6	31803040	31803102	SNORD48	11	+
chr3	186502585	186502654	SNORD2	10	+
chr17	7478031	7478165	SNORA48	10	+
chr12	57037464	57037538	SNORD59B	10	-
chr3	52726752	52726828	SNORD69	10	+
chr2	86362993	86363129	SNORD94	10	+
chr1	93302846	93302940	SNORD21	9	+
chr11	62621135	62621204	SNORD30	9	-
chr17	27050448	27050509	SNORD42A	9	+
chr1	45243514	45243584	SNORD38A	9	+
chr17	37009116	37009248	SNORA21	9	-
chr5	180670314	180670376	SNORD95	9	-

chr1	173836017	173836076	SNORD75	9	-
chr19	10218327	10218411	SNORD105	9	+
chr1	28835070	28835274	SNORA73B	8	+
chr19	17973397	17973529	SNORA68	8	+
chr17	7481273	7481409	SNORA67	8	+
chr6	133137941	133138016	SNORD100	8	+
chr15	66795583	66795652	SNORD18A	8	-
chr16	70571908	70572001	SNORD111	8	+
chr11	62623037	62623103	SNORD25	7	-
chr20	47895482	47895560	SNORD12C	7	+
chr16	2205024	2205106	SNORD60	7	-
chr17	7480129	7480276	SNORD10	7	+
chr12	49048165	49048301	SNORA34	7	-
chr7	22896232	22896305	SNORD93	7	+
chr11	62621376	62621440	SNORD29	6	-
chr18	47017653	47017717	SNORD58A	6	-
chr6	31804853	31804916	SNORD52	6	+
chr1	45244062	45244130	SNORD38B	6	+
chr5	180668818	180668889	SNORD96A	6	-
chr16	58593700	58593835	SNORA50	6	-
chr8	33370993	33371096	SNORD13	6	+
chr2	203141154	203141241	SNORD70	6	+
chr1	173835773	173835853	SNORD76	6	-
chr19	51305582	51305678	SNORD88C	6	-
chr2	101889398	101889511	SNORD89	6	-
chr19	10220425	10220516	SNORD105B	6	+

REFERENCES

Arigo, J.T., Carroll, K.L., Ames, J.M., and Corden, J.L. (2006a). Regulation of yeast NRD1 expression by premature transcription termination. Mol Cell 21, 641-651.

Arigo, J.T., Eyler, D.E., Carroll, K.L., and Corden, J.L. (2006b). Termination of cryptic unstable transcripts is directed by yeast RNA-binding proteins Nrd1 and Nab3. Mol Cell 23, 841-851.

Berndt, H., Harnisch, C., Rammelt, C., Stohr, N., Zirkel, A., Dohm, J.C., Himmelbauer, H., Tavanez, J.P., Huttelmaier, S., and Wahle, E. (2012). Maturation of mammalian H/ACA box snoRNAs: PAPD5-dependent adenylation and PARN-dependent trimming. RNA 18, 958-972.

Buratowski, S. (2009). Progression through the RNA polymerase II CTD cycle. Mol Cell 36, 541-546.

Carroll, K.L., Ghirlando, R., Ames, J.M., and Corden, J.L. (2007). Interaction of yeast RNA-binding proteins Nrd1 and Nab3 with RNA polymerase II terminator elements. RNA 13, 361-373.

Chapman, R.D., Heidemann, M., Albert, T.K., Mailhammer, R., Flatley, A., Meisterernst, M., Kremmer, E., and Eick, D. (2007). Transcribing RNA polymerase II is phosphorylated at CTD residue serine-7. Science 318, 1780-1782.

Creamer, T.J., Darby, M.M., Jamonnak, N., Schaughency, P, Hao, H., Wheelan, S.J., Corden, J.L. (2011). Transcriptome-wide binding sites for componenets of the Saccharomyces cerevisiae non-poly(A) termination pathway: Nrd1, Nab3 and Sen1. PLoS Genetics 7, e1002329.

Friedersdorf, M.B., and Keene, J.D. (2014). Advancing the functional utility of PAR-CLIP by quantifying background binding to mRNAs and lncRNAs. Genome Biol 15, R2.

Grzechnik, P., and Kufel, J. (2008). Polyadenylation linked to transcription termination directs the processing of snoRNA precursors in yeast. Mol Cell 32, 247-258.

Hirose, T., Shu, M.D., and Steitz, J.A. (2003). Splicing-dependent and -independent modes of assembly for intron-encoded box C/D snoRNPs in mammalian cells. Mol Cell 12, 113-123.

Kim, M., Vasiljeva, L., Rando, O.J., Zhelkovsky, A., Moore, C., and Buratowski, S. (2006). Distinct pathways for snoRNA and mRNA termination. Mol Cell 24, 723-734.

Lunde, B.M., Reichow, S.L., Kim, M., Suh, H., Leeper, T.C., Yang, F., Mutschler, H., Buratowski, S., Meinhart, A., and Varani, G. (2010). Cooperative interaction of

transcription termination factors with the RNA polymerase II C-terminal domain. Nat Struct Mol Biol 17, 1195-1201.

Maclary, E., Hinten, M., Harris, C., and Kalantry, S. (2013). Long nonoding RNAs in the X-inactivation center. Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology 21, 601-614.

Nedea, E., He, X., Kim, M., Pootoolal, J., Zhong, G., Canadien, V., Hughes, T., Buratowski, S., Moore, C.L., and Greenblatt, J. (2003). Organization and function of APT, a subcomplex of the yeast cleavage and polyadenylation factor involved in the formation of mRNA and small nucleolar RNA 3'-ends. J Biol Chem 278, 33000-33010.

Schaughency, P., Merran, J., Corden J.L. (2014). Genome-wide mapping of Yeast RNA Polymerase II Termination PLoS Genetics In Print.

Steinmetz, E.J., Conrad, N.K., Brow, D.A., and Corden, J.L. (2001). RNA-binding protein Nrd1 directs poly(A)-independent 3'-end formation of RNA polymerase II transcripts. Nature 413, 327-331.

Thiebaut, M., Kisseleva-Romanova, E., Rougemaille, M., Boulay, J., and Libri, D. (2006). Transcription termination and nuclear degradation of cryptic unstable transcripts: a role for the nrd1-nab3 pathway in genome surveillance. Mol Cell 23, 853-864.

Vasiljeva, L., and Buratowski, S. (2006). Nrd1 interacts with the nuclear exosome for 3' processing of RNA polymerase II transcripts. Mol Cell 21, 239-248.

Yuryev, A., Patturajan, M., Litingtung, Y., Joshi, R.V., Gentile, C., Gebara, M., and Corden, J.L. (1996). The C-terminal domain of the largest subunit of RNA polymerase II interacts with a novel set of serine/arginine-rich proteins. Proc Natl Acad Sci U S A 93, 6975-6980.

Zhang, J., and Corden, J.L. (1991). Phosphorylation causes a conformational change in the carboxyl-terminal domain of the mouse RNA polymerase II largest subunit. J Biol Chem 266, 2297-2302.

Part 3: Analyses of Nrd1-Nab3-Sen1 Termination Pathway PAR-CLIP Data Sets

Paul Schaughency and Jeffry L. Corden

INTRODUCTION

Loosely conserved regions at the 3' end of yeast genes contain signals for transcriptional termination. These termination regions (or terminators) bind proteins that are part of complexes associated with the elongating RNA Polymerase II. In protein-coding genes, the CF1 complex binds to these terminators leading to assembly of the cleavage and polyadenylation machinery (Kloetgen et al., 2014). Alternatively, Nrd1 and Nab3 recognize elements downstream of snoRNAs and cryptic unstable transcripts (CUTs) (Carroll et al., 2007; Carroll et al., 2004; Creamer, 2011; Steinmetz et al., 2001; Steinmetz et al., 2006; Wlotzka et al., 2011) and this leads to the association of a complex that contains the DNA/RNA helicase Sen1 (Vasiljeva et al., 2008). Neither of these mechanisms of termination has yet to be elucidated.

Methods of studying protein-nucleic acid interactions have relied heavily on crosslinking. Chromatin immunoprecipitation (CHiP) is commonly used to study DNA-protein interactions (Barski et al., 2007). The method uses the reversible crosslinking reagent formadehyde to purify protein-nucleic acid complexes. For RNA-protein interactions, formaldehyde crosslinking leads to high background, because of the non-specific nature of the chemical reaction. Crosslinking and immunoprecipitation (CLIP) uses UV-B to successfully and specifically crosslink protein-RNA complexes. Marrying the now inexpensive high throughput sequencing to the CLIP method (HITS-CLIP) allows for insights into the functions of many different RNA binding proteins (Darnell, 2010).

To study the Nrd1-Nab3-Sen1 (NNS) termination pathway in yeast, we have relied on the Photoactivatable Ribonucleoside enhanced Crosslinking and Immunoprecipitation (PAR-CLIP) method (Hafner et al., 2010). This has allowed us to further understand the relationship between RNA polymerase II (Pol II) and the non-poly(A) mediated termination pathway. PAR-CLIP has been used to map all binding sites of all three factors across the genome and visualize Pol II readthrough after Nrd1, Sen1, and Ysh1 mutation (Creamer, 2011; Jamonnak, 2011) .

Like HITS-CLIP, PAR-CLIP is a method to specifically crosslink RNA Binding Proteins to the RNA (Hafner et al., 2010). It does this through irradiation with 365nm UV to a culture of cells that have been treated with 4-thiouracil (4TU) or 4-thiouradine (4SU). In yeast, 4TU is easily taken up, converted to UTP and incorporated into total RNA. In mammalian cell culture, 4SU is used because of a lack of the correct uracil phosphoribosyltransferase (UPRT) to take it up 4TU from the media (Melvin et al., 1978; Miller et al., 2009). After incorporation, cells are irradiated with 365nm UV to crosslink the thio-residues that have now been included in a vast number of transcripts within the transcriptome. This wavelength of UV does not crosslink regular RNA-protein complexes. This reduces the background seen in CLIP or HITS-CLIP protocols.

Unlike formadehyde, the thio crosslink is irreversible and causes trouble for the reverse transcriptase. During the reverse transcriptase reaction, the 4TU or 4SU will cause a T to C transition in the resulting cDNA (Hafner et al., 2010). Although, harder to deal with in the analyses of these sequences, this provides a map of potential binding sites on all of the RNAs that were purified in this method.

After sequencing, PAR-CLIP has a few computational challenges that have to be overcome: Normalization of the reads from different sequencing runs and library preparations, visual representation of the sequences after mapping, and incorporating T to C transitions into the analysis. We have developed bioinformatic pipelines that try to deal with these challenges and lay a general framework for analyses of PAR-CLIP datasets in the future.

METHODS

The first pipeline is used to describe the binding of Nrd1, Nab3, and Sen1 within the transcriptome. The second pipeline was used to describe the position of Pol II after mutation of Nrd1, Sen1, or Ysh1. The third pipeline was used to describe the binding of RBM16 (a Nrd1 homolog) to RNAs in HEK-293 cells. The steps used in each pipeline are described below.

Pipeline 1 (Figure 19) : Nrd1, Nab3 , Sen1-HTB tagged datasets (Creamer, 2011; Jamonnak, 2011)

1. Program: removelinkers.R

   Author: Sarah Wheelan

   Overview:

   This program reads a fasta or fastq file and implements a dynamic matching algorithm to locate the adaptors/linkers and split them from the rest of the sequence.

   Usage:

   removelinkers.R -i input_file -o output_file_prefix -q fastq (default: T) -l1 left_linker_outer -l2 left_linker_inner -r1 right_linker_outer -r2 right_linker_inner -quart quartile, from 1-4 -int print_intermediates -el1 error_rate_start_l1 (default:0.1) -el2 error_rate_start_l2 (default:0.1) -er1 error_rate_start_r1 (default:0.1) -er2 error_rate_start_r2 (default:0.1)

   linker schema (XX is non-adaptor sequence):

l1111111-l2l2l2l2-XXXXXXXXXXXXXXX-r2r2r2r2r2-r1r1r1r1r1

Note: If there is only one right or left linker, choose either the inner or outer as they are equivalent

Where:

input_file : fasta or fastq sequence file

output_file : prefix for the files of trimmed sequences

fastq : T=fastq file, F=fasta file

left_linker_outer : 5'-most linker

left_liner_inner : second 5' linker inside outer linker

right_linker_outer : 3'-most linker

right_linker_inner : second 3' linker inside outer linker

quartile : which quartile of the file to work on, to break up big files; no entry means do the whole file

print_intermediates : should all intermediate trimmed reads be saved in separate files, or only the end product? (default = T)

error_rate_start_l1 : starting value for the dynamic error rate for l1 linker (default = 0.1)

error_rate_start_l2 : starting value for the dynamic error rate for l2 linker (default = 0.1)

error_rate_start_r1 : starting value for the dynamic error rate for r1 linker (default = 0.1)

error_rate_start_r2 : starting value for the dynamic error rate for r2 linker (default = 0.1)

Example:

removelinkers.R –i sequences.fastq –o trimmedsequences_1_ –r2

XXXXXXXXXXXXX(3’ adaptor sequence) –quart 1

2. Program: rmshortseqraw13.pl

   Author: Paul Schaughency

   Overview:

   This program reads a raw sequence file and prints out only those sequences that are longer than 13 nucleotides long.

   Usage:

   rmshortseqraw13.pl –i input_file –o output_file

   Example:

   rmshortseqraw13.pl –i sequences.raw –o sequences>13.raw

3. Program: Bowtie

   Author: (Langmead et al., 2009)

   Overview:

   A short read alignment program that uses the Burrows-Wheeler transform to efficiently map short reads to any genome.

   Usage:

   See (Langmead et al., 2009)

Example:

For 4a,b,c: bowtie –v 1 SacCer2 –m 1 –sam—nohead

For 4d: bowtie –v1 SacCer2 –m 1

4a. Program: SCfixwig.pl

Authors: Paul Schaughency, Tyler Creamer

Overview: Takes a sam output file from bowtie and converts it into a wig file used by visualization programs. It does not multiply reads to normalize them. Instead it counts 1 read as 1 read.

Usage:

SCfixwig.pl -i input sam file -p output plus wig file name -m output minus wig file name

Example:

SCfixwig.pl -i alignedsequences.sam -p treatmentplusstrand.wig -m treatmentminusstrand.wig

4b. Program: SCfixwigwithmulti.pl

Authors: Paul Schaughency, Tyler Creamer

Overview: Takes a sam output file from bowtie and converts it into a wig file used by visualization programs. It normalizes all reads to 10^7 reads by calculating

first a multiplier to apply to all reads then counting each read as a fraction of that multiplier. The multiplier is defined as the number of reads aligned/10^7 .

Usage:

SCfixwigwithmulti.pl -i input sam file -p output plus wig file name -m output minus wig file name

Example:

SCfixwigwithmulti.pl -i alignedsequences.sam -p treatmentplusstrand_norm.wig -m treatmentminusstrand_norm.wig

4c. Program: extractsamSC.pl

Author: Paul Schaughency

Overview: Splits a samfile into two files: one for the plus strand reads and one for the minus strand reads. Keeps the sam format for both files.

Usage:

extractsamSC.pl < Any Sam File

Example:

extractsamSC.pl < Alignedreads.sam

4d. Program: SCT2C.pl

Author: Paul Schaughency

Overview: Takes a raw bowtie output file and converts it into a wig file used by visualization programs. This program only counts T to C transitions.

Usage:

SCT2C.pl -i rawbowtie file -p output plus wig file name -m output minus wig file name

Example:

SCT2C.pl -i alignedsequences.rawbowtie -p treatmentplusstrand_T2C.wig -m treatmentminusstrand_T2C.wig

5. Program: binmappingv8.pl

   Author: Paul Schaughency

   Overview: Using a known annotation file and a two samfiles of plus and minus strand reads, can bin reads into any number of bins to figure out the distribution of reads within a given annotation.

   Usage:

   binmappingv8.pl [Gene Sorting Options] -g Custom Gene List -p input sam file -m input sam file -o output name

   [Gene Sorting Options]

-h sort and check for overlap in genes that have high transcription (txn > or = 6)

-d sort and check for overlap in genes that have mid range transcription (txn < 6 > 3.3)

-l sort and check for overlap in genes that have low transcription (txn < or = 3.3)

-a sort and check for overlap in all genes

-c sort and check for overlap in a custom gene list (if this is defined you must also define -g)

[Input Files]

-g Custom Gene List

Note: Must be a text file the following format:

Gene_Name <tab> Chromosome <tab> Gene Start <tab> Gene End

-p Sorted Plus Strand Sam File

-m Sorted Minus Strand Sam File

-o Base Name for all Output File

Example:

binmappingv8.pl –c –g geneannotation.txt –p plusstrandalignments.sam –m minusstrandalignments.sam –o output.txt

Pipeline 2 (Figure 20) : Nrd1 Anchor Away, Rpb2-HTB datasets (Schaughency, 2014)

1&2. See Pipeline 1

3. Program: collapserawreads.pl

   Author: Paul Schaughency

   Overview: Takes a raw bowtie output file and collapses all reads that are exactly the same into only one read.

   Usage:

   collapserawreads.pl –i input_file –o output_file

   Example:

   collapserawreads.pl –i rawsequences.raw –o collapsedrawsequences.raw

4. Program: Bowtie

   Note: See pipeline 1 step 3

   Example:

   For 5a,b: bowtie –v 2 SacCer3tRNA –m 1 –sam—nohead –best –strata -y

   For 5c: bowtie –v 2 SacCer3tRNA –m 1 –best –strata –y

5. Program: AlignedMinusExtract.pl

Overview: Converts the unread reads from bowtie and puts them back into a format so they can be used again to align to the genome.

Usage:

AlignedMinusExtract.pl –i input_file –o output_file

Example:

AlignedMinusExtract.pl –i tRNAunmappedreads.raw –o tRNAunmappedfinal.raw

6. Program: Bowtie

Note: See pipeline 1 step 3

Example:

For 5a,b: bowtie –v 2 SacCer3 –m 1 –sam—nohead –best –strata -y

For 5c: bowtie –v 2 SacCer3 –m 1 –best –strata –y

7a. Program: SCFullReadWigNoMulti.pl

Authors: Paul Schaughency

Overview: Takes a sam output file from bowtie and converts it into a wig file used by visualization programs. It does not multiply reads to normalize them. Instead it counts 1 read as 1 read. Every point across the entire read is counted.

Usage:

SCFullReadWigNoMulti.pl -i input sam file -p output plus wig file name -m output minus wig file name

Example:

SCFullReadWigNoMulti.pl -i alignedsequences.sam -p treatmentplusstrand.wig -m treatmentminusstrand.wig

7b. Program: SCFullReadWigWithMulti.pl

Authors: Paul Schaughency

Overview: Takes a sam output file from bowtie and converts it into a wig file used by visualization programs. It normalizes all reads to 10^7 reads by calculating first a multiplier to apply to all reads then counting each read as a fraction of that multiplier. Every point across the entire read is counted.

Usage:

SCFullReadWigWithMulti.pl -i input sam file -p output plus wig file name -m output minus wig file name

Example:

SCFullReadWigWithMulti.pl -i alignedsequences.sam -p treatmentplusstrand_norm.wig -m treatmentminusstrand_norm.wig

7c. Program: SCT2CV2.pl

Author: Paul Schaughency

Overview: Takes a raw bowtie output file and converts it into a wig file used by visualization programs. This program only counts T to C transitions and counts multiple T to C transitions in one read as separate entities

Usage:

SCT2CV2.pl -i rawbowtie file -p output plus wig file name -m output minus wig file name

Example:

SCT2CV2.pl -i alignedsequences.rawbowtie -p treatmentplusstrand_T2C.wig -m treatmentminusstrand_T2C.wig

8. Program: wig2bedv2.pl

   Author: Paul Schaughency

   Overview: Converts wig file format to bed file format.

   Usage: wig2bedv2.pl –i input_file –o output_file

   Example: wig2bedv2.pl –i alignedsequences.wig –o alignedsequences.bed

9. Program: globalreadthroughv3.pl

   Author: Paul Schaughency

   Overview: Calculates the global readthrough percentage for each base in the entire genome. Outputs percentage wig and bed files.

Usage: globalreadthroughv3.pl –a input bed file 1 –b input bed file 2 –c input bed file 3 –d input bed file 4 –x nucleotides upstream –y nucleotides downstream –z read cutoff

Example: Usage: globalreadthroughv3.pl –a plusstrandcontrol.bed –b minusstrandcontrol.bed –c plusstrandtreatment.bed –d minusstrandtreatment.bed –x 500 –y 500 –z 1000

Pipeline 3 (Figure 21) : RBM16-HTB Dataset

1-3. See Pipeline 2

4. Program: Bowtie

   Note: See pipeline 1 step 3

   Example:

   For 5a: bowtie –v 2 hg19tRNAs –m 1 –sam—nohead –best –strata -y

   For 5b: bowtie –v 2 hg19tRNAs –m 1 –best –strata –y

5. Program: AlignedMinusExtract.pl

   Note: See pipeline 2 step 5

6. Program: Bowtie

   Note: See pipeline 1 step 3

   Example:

For 5a: bowtie –v 2 hg19 –m 1 –sam—nohead –best –strata -y

For 5b: bowtie –v 2 hg19 –m 1 –best –strata –y

7a. Program: HGFullReadWig.pl

Authors: Paul Schaughency

Overview: Takes a sam output file from bowtie and converts it into a wig file used by visualization programs. It does not multiply reads to normalize them. Instead it counts 1 read as 1 read. Every point across the entire read is counted.

Usage:

HGFullReadWig.pl -i input sam file -p output plus wig file name -m output minus wig file name

Example:

HGFullReadWig.pl -i alignedsequences.sam -p treatmentplusstrand.wig -m treatmentminusstrand.wig

7b. Program: HGT2CV2.pl

Author: Paul Schaughency

Overview: Takes a raw bowtie output file and converts it into a wig file used by visualization programs. This program only counts T to C transitions and counts multiple T to C transitions in one read as separate entities

Usage:

HGT2CV2.pl -i rawbowtie file -p output plus wig file name -m output minus wig file name

Example:

HGT2CV2.pl -i alignedsequences.rawbowtie -p treatmentplusstrand_T2C.wig -m treatmentminusstrand_T2C.wig

DISCUSSION

The goal with any analyses like these is to try to accurately map all RNAs bound to these RNA binding factors. As described previously, PAR-CLIP was used to make cDNA libraries from factors tagged for highly denaturing tandem purification. The results of the sequencing of these libraries were large files containing millions of reads all 32bp – 50 bp long.

The first step in analyzing any library is to try to remove any sequences pertaining to the 3' adaptor. When libraries are made, adaptors are ligated at the 5' and 3' end of the RNA of interest. These adaptors allow for the reverse transcription, PCR amplification, and sequencing of these RNAs. The Illumina sequencing primer ends at the end of the end of the 5' adaptor. Therefore, none of the 5' adaptor should be present in the final sequencing reads. The 3' adaptor, although, will be present because we size selected for small insert sizes.

There are many ways to go about trimming the 3' adaptor from the sequencing reads. The most basic idea is to search for any part of the 3' adaptor and to eliminate it from the sequences. This is quite effective, but because the accuracy of the Illumina machine to incorporate the correct base decreases from 5' to 3', many reads may be thrown out accidentally. An open-source R package called Bioconductor (Gentleman et al., 2004) has been developed and contains a slightly better algorithm for 3' adaptor removal. In brief, it uses a heuristic to allow for mismatches closer to the 3' end of the read. Dr. Sarah Wheelan wrote an easy to use interface for this program, which we used in all for all of our datasets.

After trimming the sequences, depending on the amount of sequences that are trimmed, the reads can be condensed. Condensing the reads is defined as counting all exact duplicate reads as only one read. There are two reasons that reads could be the same. First, they were amplified from the same cDNA or, alternatively, they came from two different RNAs crosslinked in the same manner. Although, the latter is certainly possible, the former is a bigger concern. The presence of the T to C transitions caused by the incorporation of 4tU as well as the RNase T1 step of the library preparation point toward there being exact duplicate reads more from PCR amplification bias.

Bowtie is an efficient short read alignment program (Langmead et al., 2009). In the first experiments with PAR-CLIP, bowtie was used to conservatively align the reads to the current yeast genome (SacCer2). These reads were allowed only one mismatched base and only one unique alignment. In subsequent experiments, the reads were allowed to contain two mismatched bases and one non-unique alignment. The main difference between the first PAR-CLIP experiments and the subsequent ones was the change from 32bp reads to 50 bp reads. Allowing for only one mismatched base was deemed too conservative because of the amount of T to C transitions that could be sequenced with a longer read. This required a rewrite of all downstream programs in order to compensate for the extra information. The switch from one unique alignment to one non-unique alignment was done to try and alleviate the concern that some reads were being discarded because they were too repetitive. It is unclear whether switching from unique to non-unique would be helpful for future datasets.

Starting with the anchor away data sets, reads that mapped to tRNAs were removed from every dataset by running bowtie against a pseudogenome of all tRNAs in

yeast (or human) and then taking those reads that do no align and using those for subsequent analyses (AlignedMinusExtract.pl). Even with highly denaturing tandem affinity purification, we have seen reads map to tRNAs in every dataset (data not shown). Whether or not this is significant has yet to be determined, but for the data sets in question this was considered background.

After alignment, all datasets need to be normalized for any usable computational algorithm (SCfixwig.pl, SCfixwigwithmulti.pl, etc.) . For this, the simplest method of normalization was chosen; to adjust all read counts to that of the total reads aligned per 10^7 reads. This generally puts all reads within the same range on a per experiment basis. One caveat is that the number of mapped reads number within an experiment should be generally close to begin with or else the algorithm tends to over estimate certain data sets over others. All of our datasets within all experiments were close enough together for this to work. Other normalization techniques, as described before, could work to alleviate this problem, but, because some depend on assumptions that changes to transcription do not happen to a vast majority of transcripts, these would not necessarily be better.

The datasets, also, needed to be converted to a format so that they could be analyzed visually as well as computationally. The file formats used was .wig and .bed. Both file formats are fairly similar with major difference being that in a .wig file the chromosome header contains the chromosome number for all data points after that header while the .bed format contains the chromosome number in each line. The only other minor difference is that the .bed file format has a column dedicated to strandedness (plus or minus strand) while the .wig file format does not. This requires .wig to have a separate file for plus (Watson) strand reads and minus (Crick) strand reads. Since the

normalization happens during conversion to a wig file, sometimes the wig files had to be converted back to .bed file for computational reasons.

In the first set of PAR-CLIP experiments, the subsequent wig files were made using the midpoint of the read rounded down. If all reads were treated this way, then it should give an accurate map of approximately where the reads were crosslinked. A more accurate, but less robust way of finding crosslink sites was to only map T to C transitions. The downside of this, although it was able to get us binding motifs, was that there were much less reads to work with when you confine it to just T to C transitions. In subsequent analyses of PAR-CLIP experiments, we opted to represent every bp of the reads that mapped. This was important in order to accurately map readthrough of Pol II.

It is from here that the analyses of the individual experiments differ in terms of computational output. Most likely, programs used in a particular dataset will not be applicable to most PAR-CLIP datasets. It is worth noting, however, that the ideas behind these programs might still be usable in a different context.

Of note from the first pipeline, are the division of genes into bins and the counting of reads into the plus and minus strand of each bin. This was done by dividing a gene into a number of user defined segments, then, annotating an array of numbers each representing 1bp on each chromosome to define each bin. This came with the caveat that no two genes could overlap in the same direction or within the 5' or 3' UTR region defined by the user. This was not a problem for the gene annotation dataset used for this project because gene overlap was rare. For larger custom gene annotation lists this could eliminate too many genes from the analysis. Once all genes had defined bins all reads

were then counted that fell within each bin for each gene. Averaging these gene bins, then, was a simple computational task.

Within the anchor away dataset analysis, the calculation of percentage readthrough was unique. Similar to calculating a traveling ratio, percentage readthrough takes the difference of the percent of reads downstream of a point within the genome. This difference comes from percentages calculated from the treatment versus the control datasets. When calculated for every base in the yeast genome, it illuminates potential readthrough sites for the Nrd1 anchor away dataset. The Nrd1 anchor away dataset, though, is unique in that the readthrough is pretty long. The percentage readthrough calculation has trouble with datasets with shorter readthroughs (Sen1, Ysh1). Optimization of the calculation of percentage readthrough could alleviate this problem. Despite this caveat, the percentage readthrough calculation was able to accurately predict termination regions when the readthrough extended sufficiently downstream (Nrd1).

PAR-CLIP is a powerful method that can elucidate many unique and interesting interactions between RNA and protein within the cell. We have successfully used this method to describe the interactions of the Nrd1-Nab3-Sen1 termination pathway and RNA Pol II. We have had to develop many different bioinformatic pipelines in order to analyze the subsequent datasets. Taken together these pipelines can act as guidance for analysis of future PAR-CLIP datasets.

Figure 19. Nrd1, Nab3, Sen1 –HTB pipeline

A schematic diagram of the bioinformatic pipeline used to analyze libraries made from purifying Nrd1, Nab3, Sen1 – HTB in yeast. Where there is an arrow, the output of the previous step becomes the input of the subsequent step.

Pipeline 1

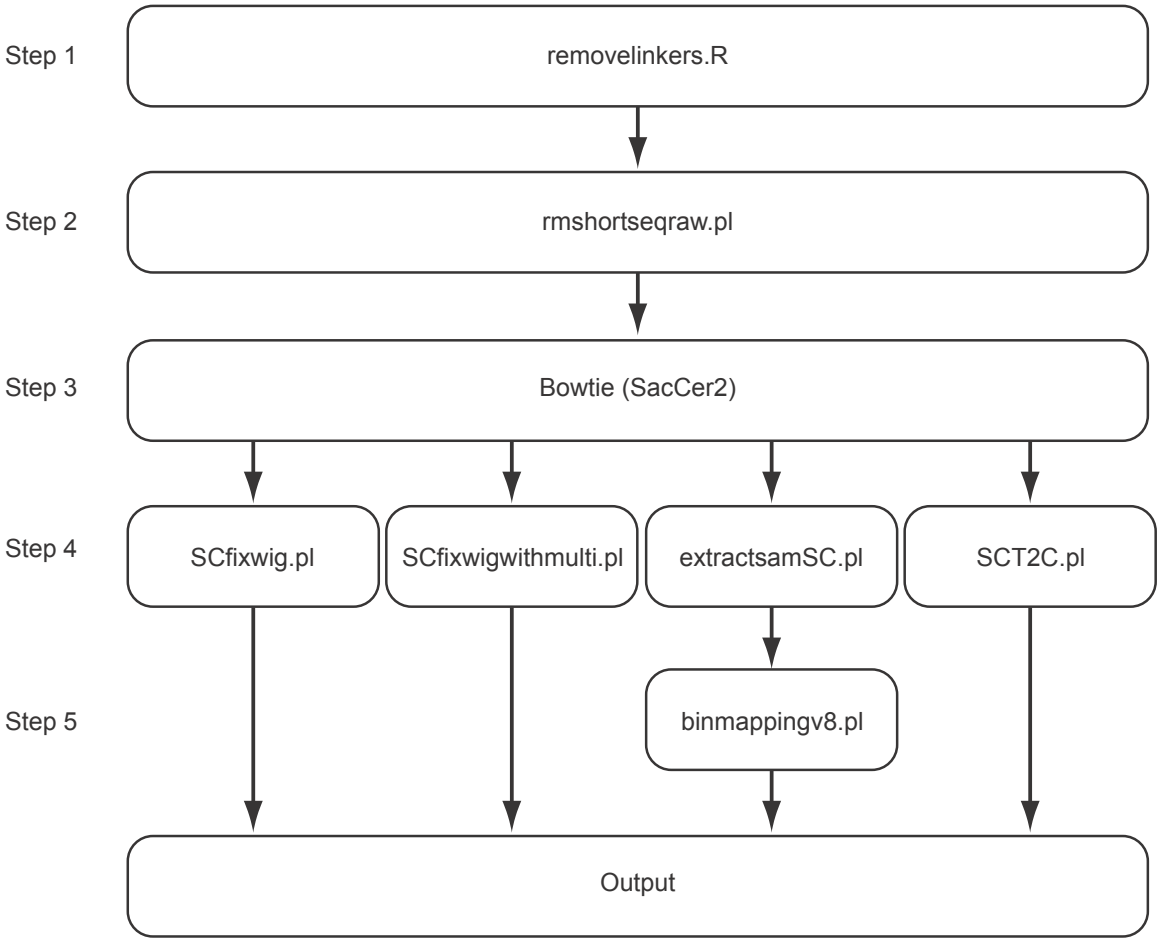

Step 1
removelinkers.R
Step 2
rmshortseqraw.pl
Step 3
Bowtie (SacCer2)
Step 4
SCfixwig.pl
SCfixwigwithmulti.pl
extractsamSC.pl
SCT2C.pl
Step 5
binmappingv8.pl
Output

Figure 20. Nrd1, Sen1, Ysh1 anchor away pipeline

A schematic diagram of the bioinformatic pipeline used to analyze libraries made from purifying Rpb2-HTB after mutation of Nrd1, Sen1, or Ysh1 in yeast. Where there is an arrow, the output of the previous step becomes the input of the subsequent step.

Pipeline 2

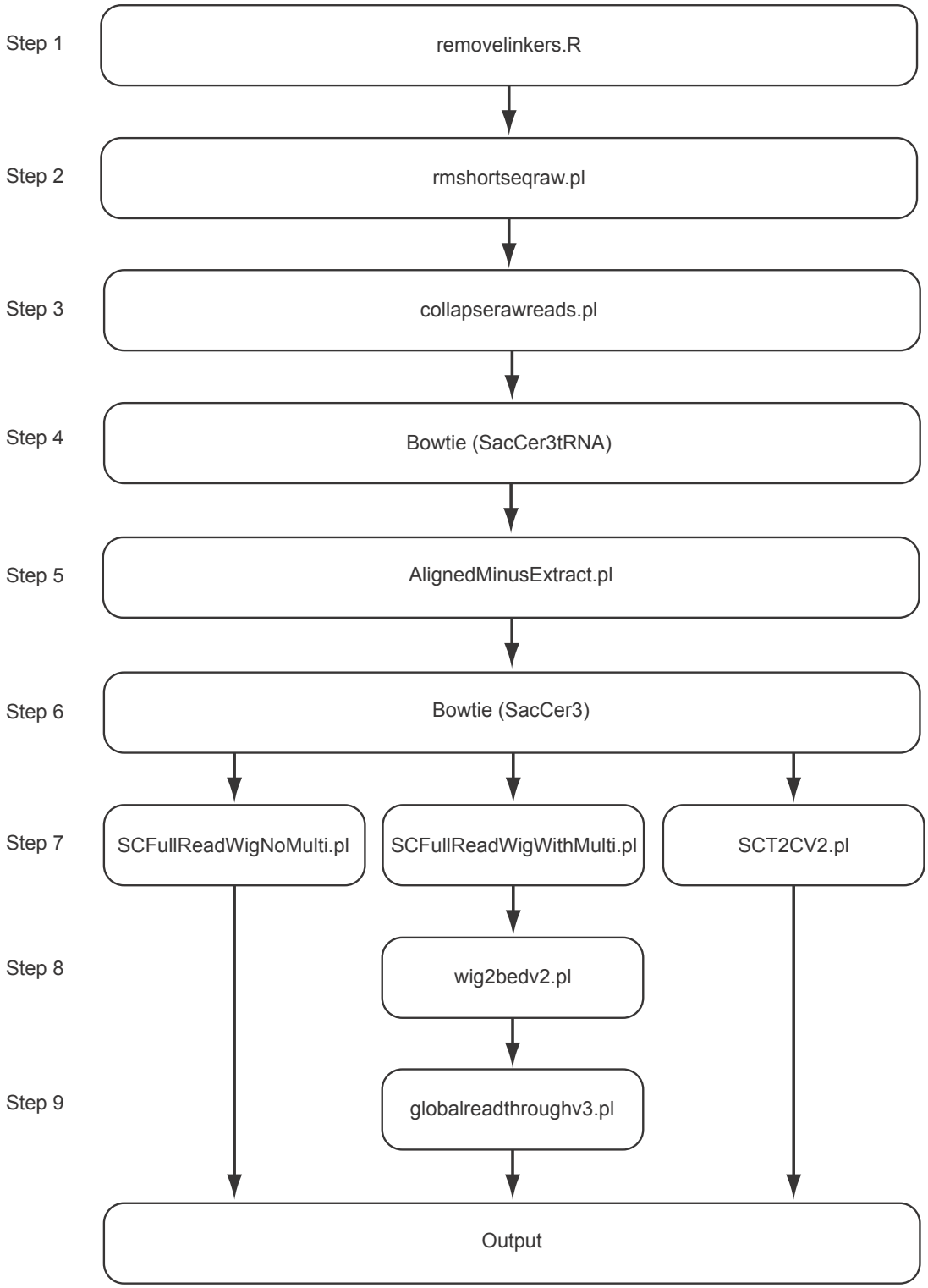
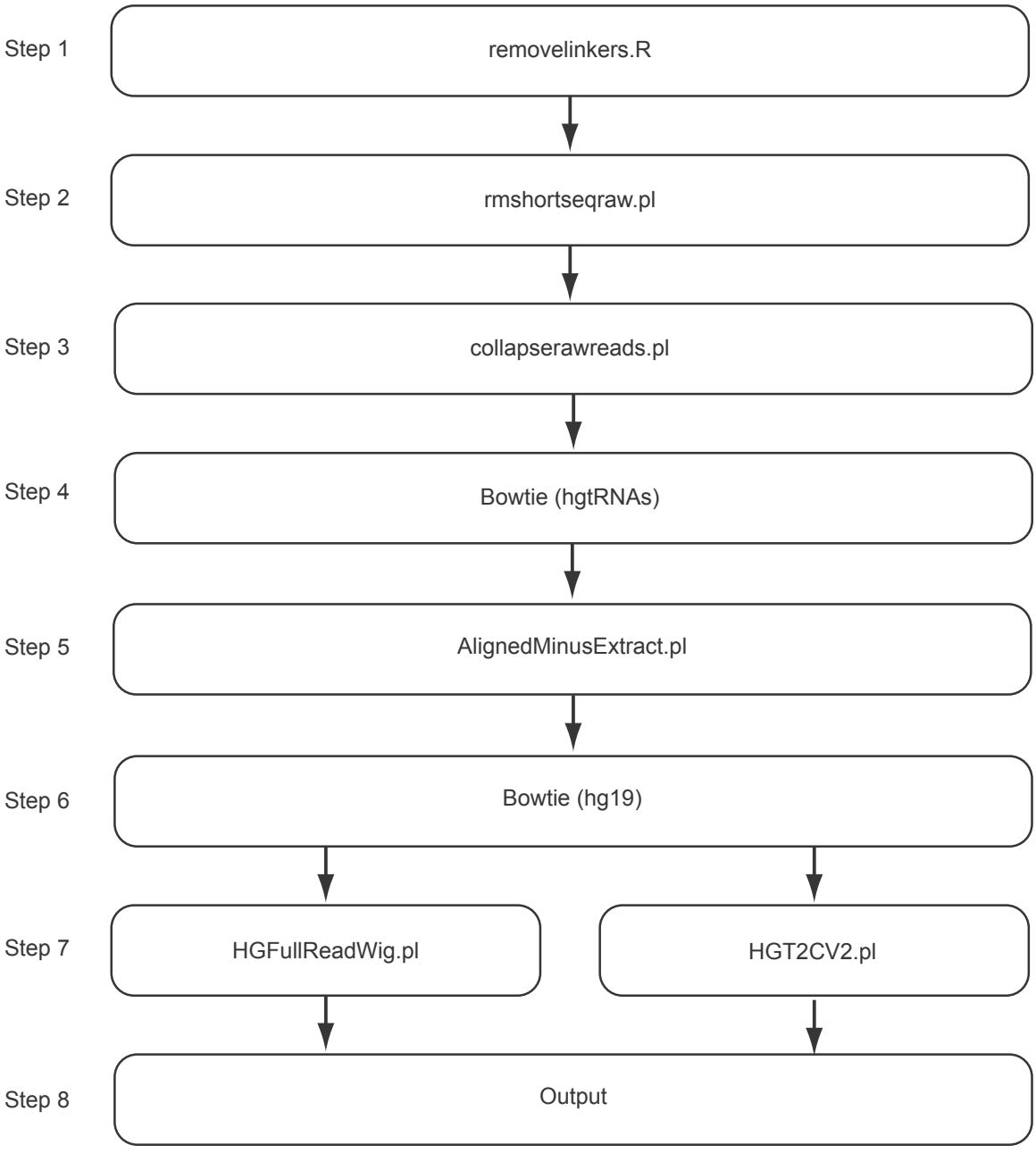

Step 1
removelinkers.R
Step 2
rmshortseqraw.pl
Step 3
collapserawreads.pl
Step 4
Bowtie (SacCer3tRNA)
Step 5
AlignedMinusExtract.pl
Step 6
Bowtie (SacCer3)
Step 7
SCFullReadWigNoMulti.pl
SCFullReadWigWithMulti.pl
SCT2CV2.pl
Step 8
wig2bedv2.pl
Step 9
globalreadthroughv3.pl
Output

Figure 21. RBM16-HTB pipeline

A schematic diagram of the bioinformatic pipeline used to analyze libraries made from purifying RBM16-HTB in HEK-293 cells. Where there is an arrow, the output of the previous step becomes the input of the subsequent step.

Pipeline 3


Step 1
removelinkers.R
Step 2
rmshortseqraw.pl
Step 3
collapserawreads.pl
Step 4
Bowtie (hgtRNAs)
Step 5
AlignedMinusExtract.pl
Step 6
Bowtie (hg19)
Step 7
HGFullReadWig.pl
HGT2CV2.pl
Step 8
Output

REFERENCES

Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. Cell *129*, 823-837.

Carroll, K.L., Ghirlando, R., Ames, J.M., and Corden, J.L. (2007). Interaction of yeast RNA-binding proteins Nrd1 and Nab3 with RNA polymerase II terminator elements. RNA *13*, 361-373.

Carroll, K.L., Pradhan, D.A., Granek, J.A., Clarke, N.D., and Corden, J.L. (2004). Identification of cis elements directing termination of yeast nonpolyadenylated snoRNA transcripts. Mol Cell Biol *24*, 6241-6252.

Creamer, T.J., Darby, M.M., Jamonnak, N., Schaughency, P, Hao, H., Wheelan, S.J., Corden, J.L. (2011). Transcriptome-wide binding sites for componenets of the Saccharomyces cerevisiae non-poly(A) termination pathway: Nrd1, Nab3 and Sen1. PLoS Genetics *7*, e1002329.

Darnell, R.B. (2010). HITS-CLIP: panoramic views of protein-RNA regulation in living cells. Wiley interdisciplinary reviews. RNA *1*, 266-286.

Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J.*, et al.* (2004). Bioconductor: open software development for computational biology and bioinformatics. Genome Biol *5*, R80.

Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano, M., Jr., Jungkamp, A.C., Munschauer, M.*, et al.* (2010). Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. Cell *141*, 129-141.

Jamonnak, N., Creamer, T., Darby, M., Schaughency, P. Wheelan, S., Corden, J. (2011). Yeast Nrd1, Nab3 and Sen1 transcriptome-wide binding maps suggest multiple roles in post-transcriptional RNA processing. RNA *17*, 2011-2025.

Kloetgen, A., Munch, P.C., Borkhardt, A., Hoell, J.I., and McHardy, A.C. (2014). Biochemical and bioinformatic methods for elucidating the role of RNA-protein interactions in posttranscriptional regulation. Briefings in functional genomics.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol *10*, R25.

Melvin, W.T., Milne, H.B., Slater, A.A., Allen, H.J., and Keir, H.M. (1978). Incorporation of 6-thioguanosine and 4-thiouridine into RNA. Application to isolation of newly synthesised RNA by affinity chromatography. European journal of biochemistry / FEBS *92*, 373-379.

Miller, M.R., Robinson, K.J., Cleary, M.D., and Doe, C.Q. (2009). TU-tagging: cell type-specific RNA isolation from intact complex tissues. Nature methods *6*, 439-441.

Schaughency, P., Merran, J., Corden J.L. (2014). Genome-wide mapping of Yeast RNA Polymerase II Termination PLoS Genetics *In Print*.

Steinmetz, E.J., Conrad, N.K., Brow, D.A., and Corden, J.L. (2001). RNA-binding protein Nrd1 directs poly(A)-independent 3'-end formation of RNA polymerase II transcripts. Nature *413*, 327-331.

Steinmetz, E.J., Ng, S.B., Cloute, J.P., and Brow, D.A. (2006). cis- and trans-Acting determinants of transcription termination by yeast RNA polymerase II. Mol Cell Biol *26*, 2688-2696.

Vasiljeva, L., Kim, M., Mutschler, H., Buratowski, S., and Meinhart, A. (2008). The Nrd1-Nab3-Sen1 termination complex interacts with the Ser5-phosphorylated RNA polymerase II C-terminal domain. Nat Struct Mol Biol *15*, 795-804.

Wlotzka, W., Kudla, G., Granneman, S., and Tollervey, D. (2011). The nuclear RNA polymerase II surveillance system targets polymerase III transcripts. EMBO J *30*, 1790-1803.

Curriculum Vitae

Paul Schaughency **October 3, 2014**

3902 Foster Ave.
Baltimore, MD 21224
pschaughency@gmail.com
4109790447

Educational History:

Ph.D. expected	2014	Program in Biochemistry, Cellular, and Molecular Biology Mentor: Dr. Jeffry Corden	Johns Hopkins School of Medicine
B.A.	2006	Biology and Biochemistry	St. Mary's College of Maryland

Publications:

Schaughency, P, Merran, J, Corden JL. **Genome-wide mapping of Yeast RNA Polymerase II Termination Sites.** In Print.

Hodge DL, Berthet C, Coppola V, Kastenmüller W, Buschman MD, Schaughency PM, Shirota H, Scarzello AJ, Subleski JJ, Anver MR, Ortaldo JR, Lin F, Reynolds DA, Sanford ME, Kaldis P, Tessarollo L, Klinman DM, Young HA. **IFN-gamma AU-rich element removal promotes chronic IFN-gamma expression and autoimmunity in mice**. J Autoimmun. 2014 Feb 28.

Ward DN, Talley DC, Tavag M, Menji S, Schaughency P, Baier A, Smith PJ. **UK-1 and structural analogs are potent inhibitors of hepatitis C virus replication.** Bioorg Med Chem Lett. 2014 Jan 15;24(2):609-12.

Creamer TJ, Darby MM, Jamonnak N, Schaughency P, Hao H, Wheelan SJ, Corden JL. **Transcriptome-wide binding sites for components of the Saccharomyces cerevisiae non-poly(A) termination pathway: Nrd1, Nab3, and Sen1**. PLoS Genet. 2011 Oct;7(10). (PMCID: PMC3197677)

Jamonnak N, Creamer TJ, Darby MM, Schaughency P, Wheelan SJ, Corden JL. **Yeast Nrd1, Nab3, and Sen1 transcriptome-wide binding maps suggest multiple roles in post-transcriptional RNA processing.** RNA. 2011 Nov;17(11):2011-25. (PMCID: PMC3198594)

Hodge DL, Yang J, Buschman MD, Schaughency PM, Dang H, Bere W, Yang Y, Savan R, Subleski JJ, Yin XM, Loughran TP Jr, Young HA. **Interleukin-15 enhances proteasomal degradation of bid in normal lymphocytes: implications for large granular lymphocyte leukemias**. Cancer Res. 2009 May 1;69(9):3986-94. (PMCID: PMC2786937)

Specialty Coursework:

Developmental Biology	2009	Johns Hopkins School of Medicine
Membrane Trafficking	2009	Johns Hopkins School of Medicine
Great Experiments in Biology	2009	Johns Hopkins School of Medicine
HIV Biology	2009	Johns Hopkins School of Medicine
Programming In Perl*	2011	Johns Hopkins School of Public Health
Introduction to Effective Instruction	2012	Johns Hopkins School of Education
Transcription Mechanisms	2012	Johns Hopkins School of Medicine
Drug Discovery Case Studies	2012	Johns Hopkins School of Medicine
Epigenetics	2013	Johns Hopkins School of Medicine

*denotes a course that was audited

Laboratory Experience:

Thesis research 2009-2014 Lab of Dr. Jeffry Corden Johns Hopkins SoM

Identified novel termination points for the Nrd1-Nab3 termination pathway genome wide using PAR-CLIP. This involved rapid conditional depletion of different termination factors from the nucleus using rapamycin followed by specific crosslinking to 4-thiouracil containing RNA bound to proteins. Sequencing was done on the purified RNA and specific bioinformatic pipelines to analyze the data were made.

Research Rotation 2008-2009 Lab of Dr. John Lorsch Johns Hopkins SoM

Examined the effects of the yeast consensus sequence on translation *in vivo* and *in vitro*

Research Rotation 2008-2009 Lab of Dr. Geraldine Seydoux Johns Hopkins SoM

Examined whether or not expression of 3'UTR GFP fusions in *c. elegans* males are viable.

CRTA Fellow 2006-2008 Lab of Dr. Howard Young NCI-Frederick

Investigated IL-15 control of bid expression in primary human natural killer cells.

Summer Internship 2005 Lab of Dr. Paul Smith UMBC

Worked on synthesizing a novel napthanol analog of UK-1 as potential Topoisomerase and HIV-1 Integrase inhibitor.

Senior Project 2004-2006 Lab of Dr. Edward Orlando St. Mary's College

Worked on treating the hermaphroditic *Rivulus marmoratus* with ethinyl estradol to try to direct the development of true females.

Presentations:

Schaughency P, Merran J, Corden J (2014) Genome-wide mapping of Yeast RNA Polymerase II Termination Sites. Molecular Biology and Genetics Departmental Research Colloquium. Johns Hopkins School of Medicine, Baltimore, MD (speaker)

Schaughency P, Corden J (2013) Genome-wide mapping of Yeast RNA Polymerase II Termination Sites. Molecular Biology and Genetics Departmental Research Colloquium. Johns Hopkins School of Medicine, Baltimore, MD (speaker)

Schaughency P (2009-2013) Presented and discussed many different interesting papers. Molecular biology and genetics journal club. Johns Hopkins School of Medicine, Baltimore, MD (speaker)

Schaughency P, Merran J, Corden J (2013) Anchoring Away Nrd1: Elucidating the Role of Non-Poly(A) Mediated Termination in Saccharomyces cerevisiae. Biochemistry, Cellular, and Molecular Biology Retreat. Johns Hopkins School of Medicine, Baltimore, MD (poster)

Schaughency P, Creamer T, Jammonak N, Darby M, Corden J (2010) RBM16: a Nrd1-like attenuator of Pol II transcription? Biochemistry, Cellular, and Molecular Biology Retreat. Johns Hopkins School of Medicine, Baltimore, MD (poster)

Schaughency P, Creamer T, Jammonak N, Darby M, Corden J (2009) RBM16 – A Novel Nrd1-like transcriptional attenuation factor in mammals? Biochemistry, Cellular, and Molecular Biology Retreat. Johns Hopkins School of Medicine, Baltimore, MD (poster)

Service and Leadership:

Student Mentor, Johns Hopkins School of Medicine. 2008-Present

Assisted Dr. Jeffry Corden with the supervision of an undergraduate student, a graduate rotation student, a technician, and an individual looking for some wet lab experience. Taught correct laboratory safety and techniques, and helped their professional development.

Biochemistry Lab Assistant, St. Mary's College of Maryland. 2005-2006

Worked with Dr. P. Mertz to prepare various solutions and gels for use in an undergraduate biochemistry laboratory. Oversaw and assisted with many biochemistry lab classes.